THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE L'UNIVERSITE DE MONTPELLIER

En Informatique

École doctorale : Information, Structures, Systèmes

Unité de recherche LIRMM

Visualisations pour la Veille en Épidémiologie Animale

Présentée par Samiha FADLOUN

Le 15 Novembre 2018

Sous la direction de Pascal PONCELET et Mathieu ROCHE

Devant le jury composé de

Julien VELCIN, Professeur, Université Lyon 2	Rapporteur
Gilles VENTURINI, Professeur, Université de Tours	Rapporteur
Carmen GERVET, Professeur, Université de Montpellier	Présidente
Renaud LANCELOT, Chercheur, CIRAD Montpellier	Examinateur
Pascal PONCELET, Professeur, Université de Montpellier	Co-directeur
Mathieu ROCHE, Chercheur, HDR, CIRAD Montpellier	Co-directeur
Arnaud SALLABERRY, Maître de Conférences, Université Paul-Valéry Montpellier 3	Co-encadrant

UNIVERSITÉ DE MONTPELLIER

Résumé

De nombreux documents concernant l'émergence, la propagation ou le suivi de maladies humaines et animales sont quotidiennement publiés sur le Web. Afin de prévenir l'expansion des maladies, les épidémiologistes doivent constamment rechercher ces documents et les étudier afin de détecter les foyers de propagation le plus tôt possible. Dans cette thèse, nous nous intéressons aux deux activités liées à ce travail de veille afin de proposer des outils visuels permettant de faciliter/accélérer l'accès aux informations pertinentes. Nous nous focalisons sur les maladies animales, qui ont été moins étudiées et qui pourtant peuvent avoir de lourdes conséquences sur les activités humaines (maladies transmises d'animaux à humains, épidémies dans les élevages, ...).

La première activité du veilleur consiste à collecter les documents issus du Web. Pour cela, nous proposons EpidVis, un outil visuel permettant aux épidémiologistes de regrouper et structurer les mots-clés nécessaires à leurs recherches, construire visuellement des requêtes complexes, les lancer sur différents moteurs de recherche et visualiser les résultats retournés. La seconde activité du veilleur consiste à explorer un grand nombre de documents concernant les maladies. Ces documents contiennent non seulement des informations telles que les noms des maladies, les symptômes associés, les espèces infectées, mais aussi des informations de type spatio-temporelles. Nous proposons EpidNews, un outil de visualisation analytique permettant d'explorer ces données en vue d'en extraire des informations. Les deux outils ont été réalisés dans le cadre d'une étroite collaboration avec des experts en épidémiologie. Ces derniers ont réalisé des études de cas pour montrer que les fonctionnalités des propositions étaient complètement adaptées et permettaient de pouvoir facilement extraire de la connaissance.

Mots clés : visualisation d'informations, visualisation analytique, veille épidémiologique animale.

Abstract

Many documents concerning emergence, spread or follow-up of human and animal diseases are published daily on the Web. In order to prevent the spread of disease, epidemiologists must frequently search for these documents and analyze them to detect outbreaks as early as possible. In this thesis, we are interested in the two activities related to this monitoring work in order to produce visual tools facilitating the access to relevant information. We focus on animal diseases, which have been less studied but can have serious consequences for human activities (diseases transmitted from animals to humans, epidemics in livestock ...).

The first activity is to collect documents from the Web. For this, we propose EpidVis, a visual tool that allows epidemiologists to group and organize the keywords used for their research, visually build complex queries, launch them on different search engines and view the results returned. The second activity is to explore a large number of documents concerning diseases. These documents contain not only information such as disease names, associated symptoms, infected species, but also spatio-temporal information. We propose EpidNews, a visual analytics tool to explore this data for information extraction. Both tools were developed in close collaboration with experts in epidemiology. The latter carried out case studies to show that the functionalities of the proposals were completely adapted and made it possible to easily extract knowledge.

Keywords: information visualization , visual analytics, animal disease surveillance.

Remerciements

Tout d'abord, je remercie le ministère de l'enseignement supérieur et de la recherche scientifique de l'Algérie pour avoir financé cette thèse. Je remercie aussi l'équipe ADVANSE, le LIRMM et l'Université de Montpellier (UM) qui m'ont bien accueilli et m'ont fourni des conditions très favorables pour réaliser ma thèse.

Je voudrais remercier profondément mes encadrants, Pascal Poncelet, Mathieu Roche et Arnaud Sallaberry pour m'avoir guidé et aidé à renforcer mes connaissances. Je les remercie aussi pour leur patience et encouragement pendant les moments durs. Je tiens à remercier Julien Rabatel pour son support et sa présence dans plusieurs réunions durant la première année de ma thèse.

Je remercie les membres de mon jury. Gilles Venturini et Julien Velcin pour avoir accepté de relire ma thèse et pour toutes les remarques pertinentes dans leurs rapports. Je remercie Carmen Gervet et Renaud Lancelot pour avoir accepté d'être mes examinateurs et d'évaluer mon travail.

Je voudrais remercier toutes les personnes des équipes ASTRE, TETIS et CIRAD qui ont contribué aux travaux de cette thèse. Je remercie aussi les épidémiologistes, Elena Arsevska, Alizé Mercier et Sarah Valentin pour tester et évaluer certains travaux dans cette thèse.

Un grand merci aux autres membres de mon équipe : Sandra, Jérôme, Nancy, Erick, Jessica, Mike, Vijay, Amine, Sarah, Bilel, Fati et Walide.

Je tiens aussi à remercier ma famille, ma mère Massika (Jamila) Hadria, mon père Sebti (Farock) Fadloun, mes deux frères Toufik et Abdelraouf et mes deux sœurs Marwa et Fatima pour leurs immenses supports et encouragement dans toute la période de ma thèse. Un grand merci à mon mari, Kheireddine, qui m'a supporté et m'a encouragé dans la dernière année de la thèse.

Pour finir, je voudrais aussi remercier mes amies qui m'ont aidé dans cette période, FUNNY GROUP : Anfel, Soumia, Sakina, Nadira, Asma, Lynda et Zineb. Grâce à elles j'ai pu supporter le fait d'être très loin de ma famille.

Sommaire

Liste des figures

Liste des tables

Chapitre

1 Introduction

Sommaire

1.1 Contexte

Ces dernières années, notamment sur le Web, on a pu observer une forte augmentation du nombre d'informations publiées concernant la détection, l'émergence, la propagation ou le suivi de maladies à travers le monde (humaines et animales). Ces informations jouent un rôle majeur pour la veille épidémiologique. Un des objectifs des épidémiologistes est de prévenir la propagation des maladies. Pour cela, ils ont besoin de construire et lancer des requêtes sur le web, puis de recueillir et d'analyser les grandes quantités de données retournées.

La propagation d'une maladie animale peut avoir de lourdes conséquences pour l'Homme. En particulier, elle peut engendrer différents types de pertes directes (e.g. animaux malades ou morts dans une exploitation agricole) ou indirectes (augmentation du coût de production, entrave aux échanges commerciaux). Les épidémies peuvent ainsi avoir de graves conséquences socio-économiques et politiques [1]. Par exemple, la *maladie de la langue bleue (bluetongue)*, peut causer la mort de *moutons*, *bovins*, *chèvres* et autres *ruminants sauvages*. De plus, les animaux peuvent être porteurs de pathogènes zoonotiques, i.e. transmissibles de l'animal à l'homme et vice-versa [69], tels que la grippe (transmise par l'air) ou la rage (transmise par une morsure ou plus généralement par la salive). Malgré des observations, la plupart des données disponibles sur le Web portent sur les maladies humaines.

Dans cette thèse, nous nous concentrons donc sur l'épidémiologie animale en proposant des outils qui aident les veilleurs de ce domaine à retrouver et explorer des informations publiées sur le Web.

1.2 Problématiques

Les problématiques que nous traitons ici sont basées sur les besoins exprimés par des épidémiologistes. Ces besoins ont été identifiés en collaboration avec eux au cours de plusieurs réunions. Dans cette section, nous allons présenter succinctement ces problématiques (nous les détaillerons dans les chapitres suivants). Ensuite, nous donnerons une introduction générale au domaine de la visualisation d'informations qui servira de point de départ à nos propositions pour les résoudre.

Construction visuelle de requêtes

La détection et le suivi de maladies animales à partir de données issues du Web constituent des activités fondamentales pour la veille épidémiologique. Concrètement, les épidémiologistes interrogent régulièrement les pages web à l'aide de diverses requêtes pour obtenir de nouvelles informations. Les mots-clés des requêtes sont habituellement stockés dans des fichiers texte non structurés et les veilleurs copient-collent ces mots dans leur moteur de recherche en essayant de les associer

1. http://agriculture.gouv.fr/maladies-animales

au mieux en fonction de leurs expériences/connaissances antérieures. Par exemple, pour suivre la maladie de la langue bleue, l'épidémiologiste va associer cette maladie avec une série de symptômes ou d'espèces liées. Cette tâche peut être pénible à réaliser et des associations peuvent être oubliées. C'est pourquoi un outil visuel d'aide permettant de structurer les mots-clés, de générer des requêtes à partir de ceux sélectionnés par l'utilisateur et de visualiser les résultats selon les différents moteurs de recherche pourrait faciliter/accélérer le travail de veille.

Outre le besoin de gérer ses propres mots-clés et leurs associations, un épidémiologiste peut aussi vouloir intégrer à ses requêtes des informations transmises par ses collègues ou extraites du Web via des outils de fouille de textes. Ces ensembles de mots-clés/associations peuvent constituer un moyen efficace pour enrichir les requêtes du veilleur et ainsi extraire des pages Web jugées plus pertinentes. Une donnée intéressante pour un épidémiologiste peut, par exemple, être un nouveau symptôme détecté pour la maladie de la langue bleue. Cependant, les informations externes peuvent aussi contenir des ensembles de données peu intéressantes selon le contexte. Un outil visuel permettant de charger les données externes, les visualiser, les comparer avec les mots-clés/associations déjà saisies par l'utilisateur, et les ajouter en fonction de leur pertinence peut ainsi s'avérer d'une grande utilité pour accélérer le travail de veille.

Visualisation des sources de données

Habituellement, les données extraites du Web et portant sur les épidémies animales contiennent à la fois des informations géographiques et des informations temporelles. Par exemple, une nouvelle (*news*) peut nous apprendre que la maladie de la langue bleue a été détectée en Islande le 10 avril 2018. Des données de ce type sont appelées informations spatio-temporelles. Elles sont particulièrement utiles aux épidémiologistes dans la mesure où elles leur permettent de trouver les zones d'émergence ou de diffusion d'une maladie, et ainsi de proposer des solutions pour éviter qu'elles se propagent plus largement.

Les données Web peuvent être réparties dans deux catégories : (1) celles provenant de sources officielles comme l'Organisation mondiale de la santé animale (OIE), et (2) celles provenant de sources non officielles, comme des dépêches. Les premières contiennent des informations vérifiées et structurées, dans lesquelles il est facile d'extraire les informations spatio-temporelles. Cependant, elles sont publiées généralement assez tardivement et c'est pourquoi les épidémiologistes s'intéressent aussi aux sources non officielles, qui sont certes de qualité moindre mais qui peuvent permettre de détecter plus rapidement certaines maladies. Dans ce cas des techniques de fouille de textes peuvent être mises en place pour en extraire les informations spatio-temporelles selon un certain degré de précision. En plus de ces informations, les sources contiennent aussi d'autres données utiles, comme les *maladies* bien sûr, mais aussi les *symptômes* observés, les *espèces* impactées, etc. Analyser de telles données complexes, surtout quand elles sont volumineuses, ne peut pas être fait

manuellement de façon efficace. C'est pourquoi les épidémiologistes ont exprimé le besoin d'un outil de visualisation permettant d'explorer les différentes dimensions des données.

Introduction à la visualisation d'informations

Comme nous l'avons vu, les problématiques mentionnées ci-dessus nécessitent la mise en place d'outils visuels.

La visualisation d'informations est un domaine étudiant les systèmes de communication d'informations basés sur des représentations graphiques éventuellement interactives. Ces graphiques peuvent représenter des grandes quantités d'informations beaucoup plus rapidement que ne le feraient des pages textuelles [95].

Le but de la visualisation est donc de produire des représentations permettant à un utilisateur d'extraire des connaissances à partir d'un ensemble de données. Pour cela, le concepteur dispose d'objets graphiques (points, lignes, surfaces) qu'il va positionner sur une, deux ou éventuellement trois dimensions. Ces objets représentent des éléments d'informations dont les attributs associés peuvent être représentés à l'aide de variables visuelles contrôlant l'aspect des objets. Par exemple, un attribut quantitatif peut être représenté à l'aide de l'aire des objets graphiques ou de leurs positions en x ou en y, un attribut catégoriel peut être représenté à l'aide de leurs teintes ou de leurs formes, un ensemble d'attributs peut être représenté par des glyphes intégrant différentes variables visuelles, etc. [17, 40, 94, 96].

Un exemple typique de visualisation permettant d'accéder à un riche ensemble de données a été proposé par Minard [91] au XVIIIe siècle. Ce graphique représente la campagne de Napoléon en Russie (Cf. Figure 1.1). Il est remarquable dans la mesure où il contient beaucoup d'informations intégrées sur une seule carte : la taille de l'armée et sa localisation au cours du temps, sa direction, l'évolution des températures. Par exemple, on repère facilement que la taille de l'armée a quasiment diminué de moitié lors du passage de la Bérézina.

La Figure 1.2 montre le processus général de la visualisation [95]. Les données se présentent dans un premier temps sous une forme non structurée. Il faut ensuite les transformer de façon à obtenir des ensembles structurés pouvant être lus/utilisés par les outils de visualisation (tables de données). L'encodage visuel consiste ensuite à définir comment les données (éléments, attributs) vont être représentées. Ceci permet d'obtenir des structures visuelles. Enfin, la vue permet à l'utilisateur de voir les structures visuelles ainsi construites.

Lorsque les données sont complexes ou volumineuses, il est souvent utile de permettre à l'utilisateur d'interagir avec les différents composants du pipeline de la visualisation. Par exemple, l'utilisateur peut vouloir modifier les tables de données en supprimant (filtrant) des attributs qui ne lui semblent pas pertinents. Il peut aussi vouloir modifier l'encodage visuel, par exemple en passant d'un type de graphique à un autre. Enfin, il peut modifier la vue, par exemple en changeant le niveau de zoom.

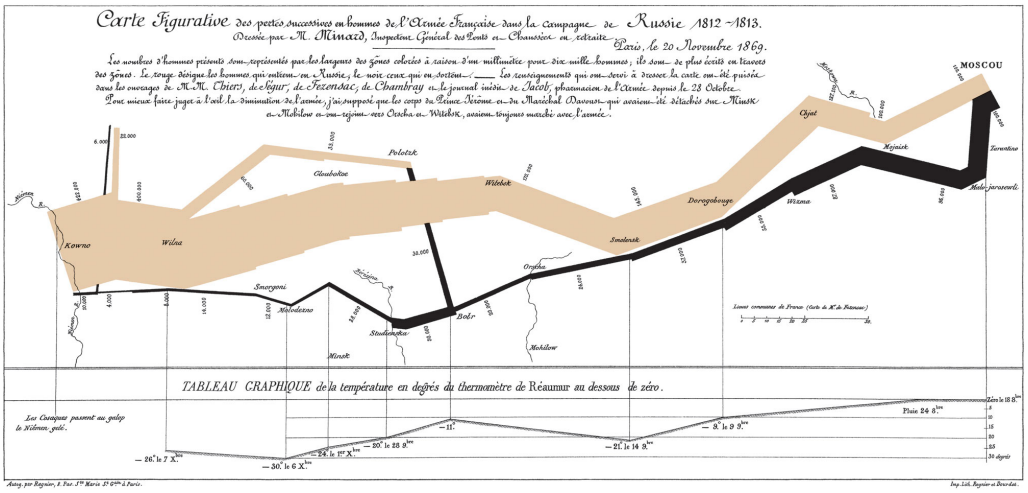


FIGURE 1.1 – La campagne de Russie de Napoléon réalisée par Minard [91]. La polyligne jaune représente la marche en direction de Moscou et la polyligne noire représente le retour. La largeur de ces polylignes correspond au nombre de soldats évoluant au cours du temps. Le diagramme en lignes en-dessous montre les températures.

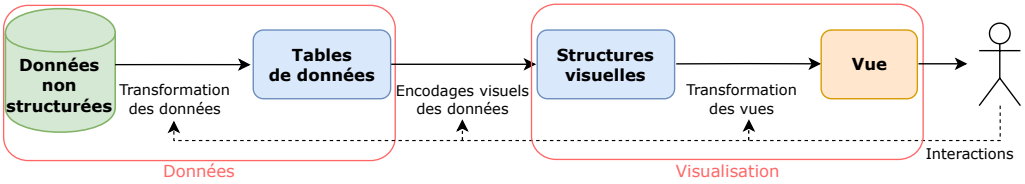


FIGURE 1.2 – Pipeline de la visualisation [95].

La visualisation analytique est la science du raisonnement analytique facilité par des interfaces visuelles interactives [25]. Elle permet de concevoir des outils facilitant la compréhension des données volumineuses et complexes, de raisonner dessus, et ainsi d'aider les utilisateurs à prendre des décisions appropriées. La recherche en visualisation analytique est interdisciplinaire et combine divers domaines de recherche complémentaires tels que la visualisation, la fouille de données, la gestion de données, la fusion de données, les statistiques et les sciences cognitives (entre autres) [55].

1.3 Contributions et organisation du mémoire

Pour répondre à la problématique concernant la création de requêtes Web, nous proposons EPIDVIS (Cf. Figure 1.3). Cet outil est composé de plusieurs vues interactives permettant de gérer l'ensemble de la chaîne de requêtage : (1) gestion des mots-clés et de leurs relations, (2) construction, affinage et lancement des requêtes,

(3) visualisation des résultats et (4) intégration de suggestions issues d'un fichier externe. La Figure 1.3 illustre les différentes vues mises en œuvre.

FIGURE 1.3 – EPIDVIS.

Lors de la conception d'EPIDVIS, un problème de chevauchements d'éléments sur une dimension s'est posé et sa résolution constitue la deuxième contribution de cette thèse. L'idée ici est de proposer un algorithme prenant en entrée une visualisation avec chevauchements et permettant de déplacer les éléments de façon à supprimer ces chevauchements. Dans un premier temps, nous proposons quatre critères définissant les propriétés que doit respecter la visualisation finale. Ensuite, nous proposons un algorithme permettant de repositionner les nœuds tout en respectant ces critères. Sa complexité temporelle est en $O(|V|log(|V|)$ (où V correspond à l'ensemble des nœuds).

Notre troisième contribution est EPIDNEWS (Cf. Figure 1.4), un système de visualisation analytique. Il répond à la deuxième problématique en permettant d'explorer des données épidémiologiques et spatio-temporelles issues de sources officielles et non officielles. Différentes vues sont mises à la disposition de l'utilisateur. Elles sont interactives et inter-connectées de façon à permettre à l'utilisateur de pouvoir filtrer les données pour se focaliser sur ses objets d'étude. La Figure 1.4 illustre les différentes vues mises en œuvre.

Le mémoire est organisé de la manière suivante. Le chapitre 2 présente EPIDVIS en décrivant les différentes vues proposées ainsi que les évaluations menées. Notre algorithme de suppression de chevauchements d'éléments en 1D est présenté dans la chapitre 3. EPIDNEWS est présenté et évalué dans le chapitre 4. Enfin, nous concluons et nous présentons les perspectives associées à nos travaux de recherche dans le chapitre 5.

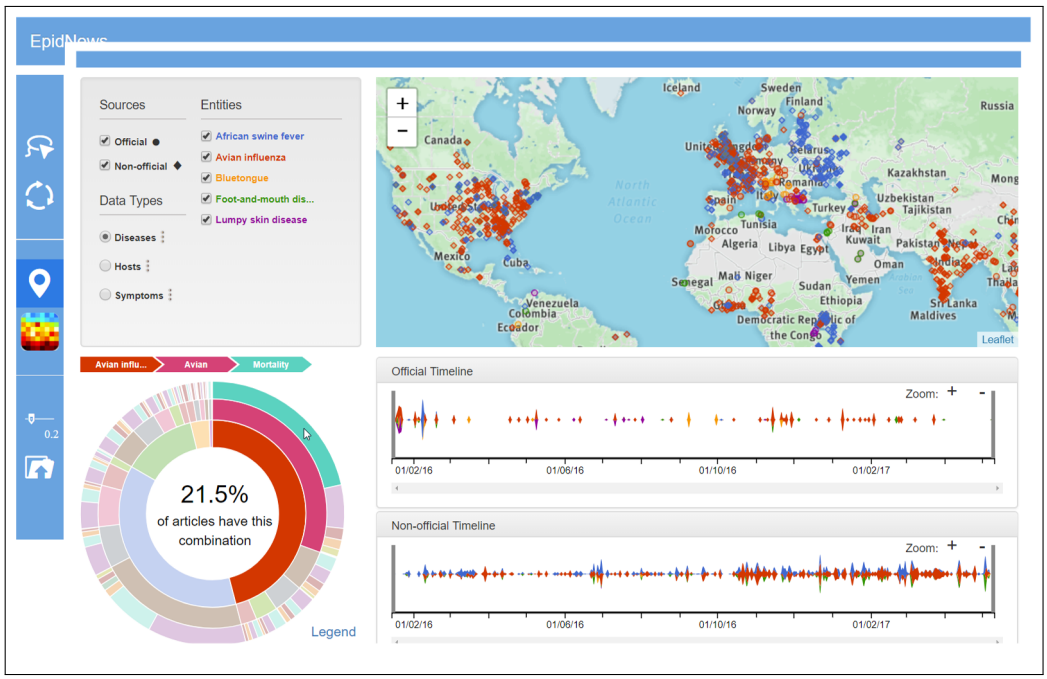


FIGURE 1.4 – EPIDNEWS.

Chapitre 2

Système de construction visuelle de requêtes pour la veille en épidémiologie animale

Sommaire

2.1 Introduction

Des organisations officielles (e.g. Organisation Mondiale de la Santé Animale (OIE)) ont pour objectif de diffuser des informations sur les épidémies. Cependant, il existe souvent un décalage entre l'apparition d'une épidémie et sa déclaration officielle. Celui-ci est généralement lié à la nécessité de nombreuses validations avant la notification officielle. Il existe, cependant, de nombreuses informations disponibles sur les épidémies animales (e.g. l'émergence d'un nouveau pathogène, l'apparition d'une nouvelle maladie) qui sont publiées en ligne via des sources non officielles (e.g. dépêches de presse). Dans le cadre de la veille, les épidémiologistes doivent donc surveiller quotidiennement de multiples sources et restent attentifs aux informations susceptibles de montrer l'apparition d'épidémies. Pour cela, ils disposent, bien entendu, de flux RSS associés à des sites spécifiques mais, la plupart du temps, ils effectuent des requêtes sur des moteurs de recherche et doivent traiter un grand nombre de résultats. L'une des premières difficultés qu'ils rencontrent est l'écriture de la requête elle-même : comment écrire une requête précise en fonction d'un moteur pour qu'il puisse retourner les résultats les plus pertinents ? Comment représenter le mieux possible les connaissances de l'expert pour en tenir compte dans la requête ? La seconde difficulté est liée au nombre de résultats retournés : comment les organiser, les sauvegarder ? Enfin une troisième difficulté concerne l'utilisation de connaissances externes issues d'autres experts ou d'approches de fouille de textes : comment les intégrer facilement pour pouvoir améliorer l'expression des requêtes ?

Dans ce chapitre, nous présentons EpidVis [36] une plateforme d'aide à la construction visuelle de requêtes pour la veille en épidémiologie animale. Elle est composée de différentes vues pour aider les utilisateurs à construire et à exécuter automatiquement les requêtes sur différents moteurs de recherche. Elle propose également des fonctionnalités pour visualiser, trier et sauvegarder les résultats de requêtes. Enfin, elle permet d'intégrer des connaissances externes afin d'offrir aux experts des suggestions de nouvelles relations et ainsi construire des requêtes plus précises.

Ce chapitre est organisé de la manière suivante : la section 2.2 présente la problématique et l'état de l'art associé. La section 2.3 décrit EpidVis avec les différentes vues associées : le gestionnaire de mots-clés, le constructeur de requêtes, la visualisation des résultats et la visualisation de suggestions. Nous proposons dans la section 2.4 trois évaluations différentes : la première est une évaluation par des utilisateurs des fonctionnalités proposées, la seconde est une étude de cas réalisée par des épidémiologistes avec EpidVis et la dernière est une discussion sur les requêtes complexes. La section 2.5 conclut ce chapitre par une discussion.

2.2 Problématique et état de l'art

De nombreuses réunions entre informaticiens et épidémiologistes de la plateforme Française de veille épidémiologique ont permis de mieux comprendre le processus de

veille et de mettre en évidence une liste de critères importants à prendre en compte dans notre application.

Les épidémiologistes disposent et utilisent de nombreuses informations issues d'organismes officiels, comme l'Organisation Mondiale de la Santé Animale (OIE) (Cf. Figure 2.1), qui leur permettent de suivre des épidémies, d'avoir des compléments d'information, de connaître des bilans, etc.

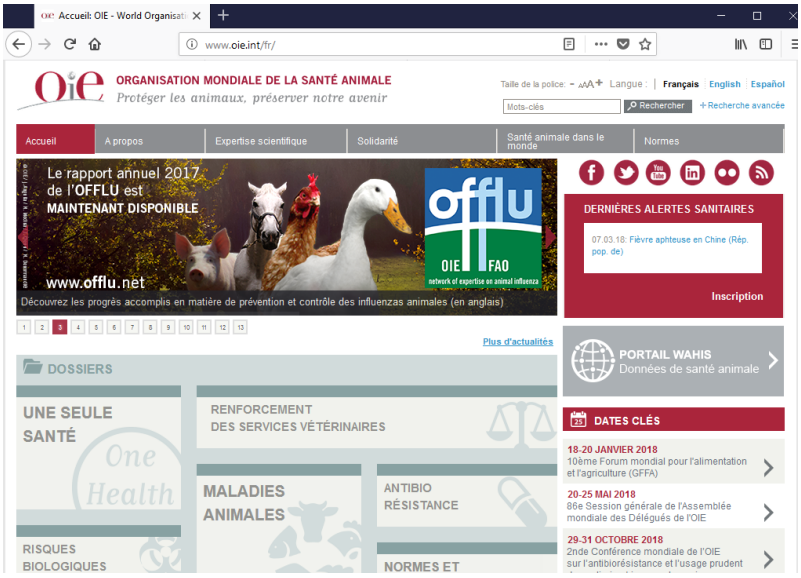
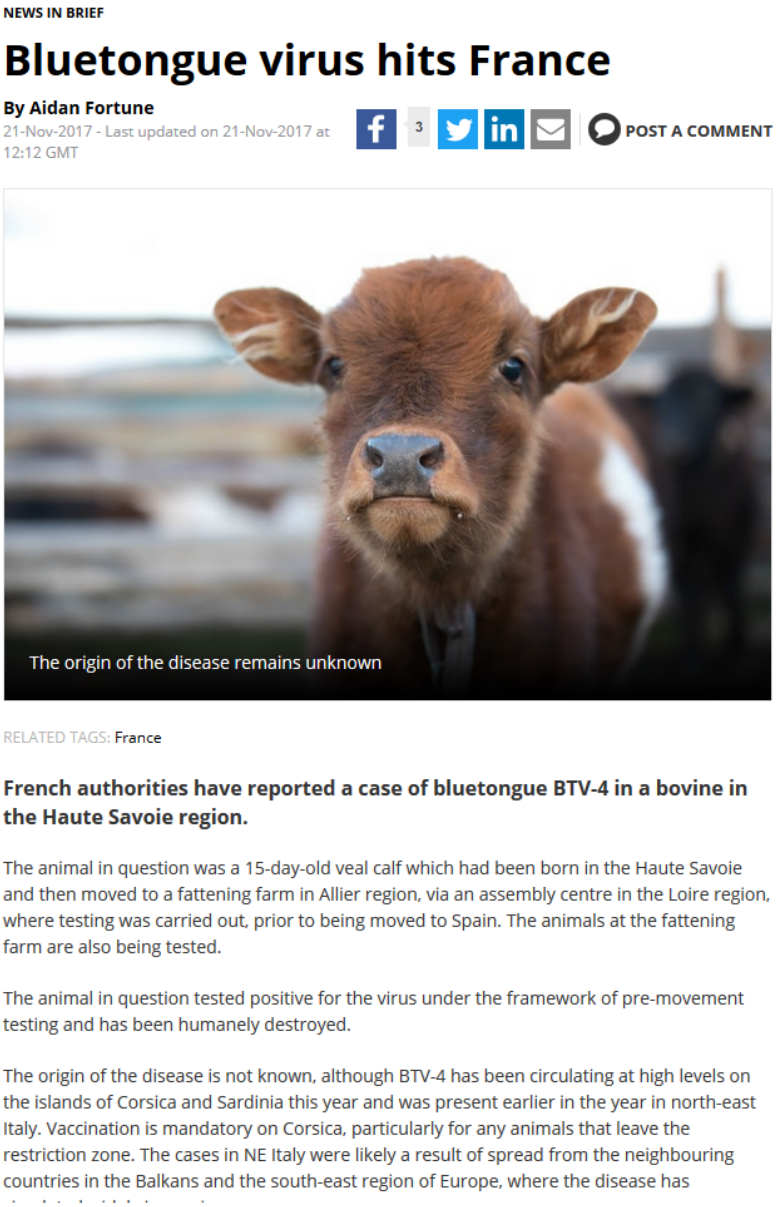


FIGURE 2.1 – Le site officiel de l'Organisation Mondiale de la Santé Animale (OIE).

Cependant ces informations officielles mettent souvent du temps à être validées et mises à disposition. Pour pallier ce problème, les utilisateurs effectuent régulièrement des recherches sur le web pour identifier des dépêches de presse afin de vérifier si de nouvelles épidémies ne sont pas en train d'apparaître. Par exemple, la Figure 2.2 illustre une dépêche montrant qu'un nouveau cas de bluetongue est apparu en France.

Dans ce contexte, trois catégories principales de termes qui sont fortement liées ont été mises en évidence : les *maladies*, les *espèces* et les *symptômes*. Par exemple, la *peste porcine africaine* est une maladie, le *cochon* est une espèce, et l'*hémorragie* est un symptôme. Pour rechercher des informations, les experts interrogent alors les moteurs de recherche via des requêtes du type *"peste porcine africaine cochon hémorragie"* et analysent les résultats obtenus comme l'illustre la Figure 2.3.

La tâche de la collecte de documents est bien entendu difficile et fastidieuse dans la mesure où elle nécessite d'analyser de très nombreux résultats et que souvent ceux-ci ne sont pas adaptés : contenus trop généralistes, informations déjà connues, etc. Le travail collaboratif avec des experts a également mis en évidence qu'il pouvait y avoir, de fortes corrélations entre les catégories. Par exemple, la *peste porcine Africaine* est très fortement liée au *cochon* mais peut également toucher les *sangliers*. A l'heure



NEWS IN BRIEF

Bluetongue virus hits France

By Aidan Fortune
21-Nov-2017 - Last updated on 21-Nov-2017 at 12:12 GMT

3 POST A COMMENT

The origin of the disease remains unknown

RELATED TAGS: France

French authorities have reported a case of bluetongue BTV-4 in a bovine in the Haute Savoie region.

The animal in question was a 15-day-old veal calf which had been born in the Haute Savoie and then moved to a fattening farm in Allier region, via an assembly centre in the Loire region, where testing was carried out, prior to being moved to Spain. The animals at the fattening farm are also being tested.

The animal in question tested positive for the virus under the framework of pre-movement testing and has been humanely destroyed.

The origin of the disease is not known, although BTV-4 has been circulating at high levels on the islands of Corsica and Sardinia this year and was present earlier in the year in north-east Italy. Vaccination is mandatory on Corsica, particularly for any animals that leave the restriction zone. The cases in NE Italy were likely a result of spread from the neighbouring countries in the Balkans and the south-east region of Europe, where the disease has

FIGURE 2.2 – Un exemple de dépêche.

actuelle de telles relations ne s'expriment que par l'intermédiaire d'ajout de mots-clés dans la requête, e.g. *"peste porcine africaine cochon sanglier"*.

Les moteurs de recherche proposent de nombreuses fonctionnalités pour pouvoir affiner les recherches : ajout d'opérateurs logiques AND, OR, NOT; sélection d'intervalles de temps, choix de moteurs plus spécialisés (GoogleNews vs. Google), etc. Ces fonctionnalités sont cependant souvent spécifiques aux différents moteurs et nécessitent un apprentissage pour les utiliser au mieux. Même si les épidémiologistes

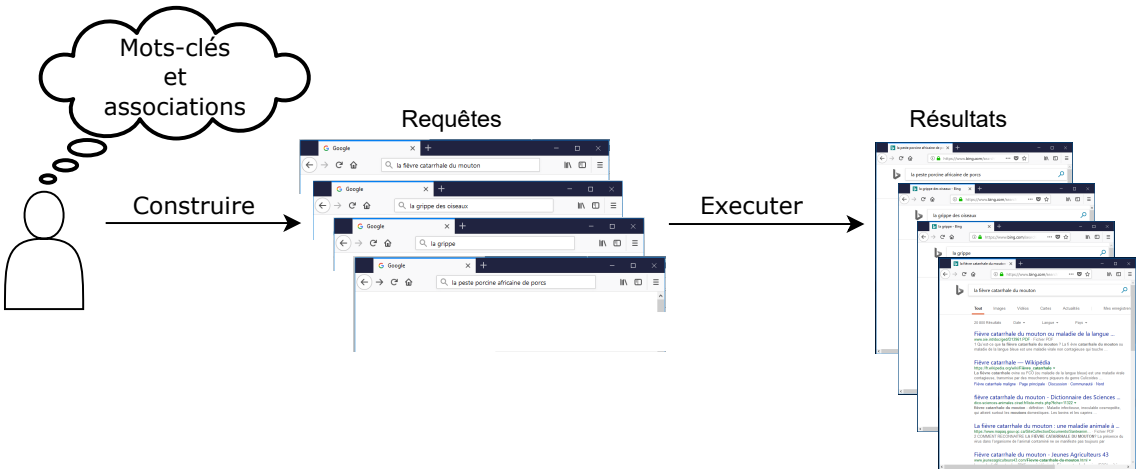


FIGURE 2.3 – Processus général de veille.

peuvent les utiliser, ils souhaiteraient que la prise en compte de ces spécificités puisse se faire de manière plus spécifique tout en garantissant une certaine "systématicité".

Ces différentes utilisations nous ont donc permis, en collaboration avec les utilisateurs, de définir des critères que devraient vérifier une approche d'aide à la création de requêtes que nous présentons dans la section suivante.

2.2.1 Les critères

L'une des manières d'améliorer le processus de veille actuel est de pouvoir répondre à la question : "est-il possible de proposer une approche pour construire facilement et dynamiquement des requêtes qui intègrent des connaissances expertes afin de rechercher de l'information pertinente ?". Pour répondre à cette question nous avons mis en évidence, en collaboration avec des épidémiologistes, les critères suivants que devrait proposer une application d'aide à la veille.

C1 : Gestion de mots-clés

Les requêtes sont construites à l'aide de combinaisons de mots-clés. L'application doit être capable de prendre en compte les trois catégories principales : maladie, espèce et symptôme. La combinaison de ces catégories doit pouvoir se faire de manière transparente pour les utilisateurs.

C2 : Gestion de relations entre les mots-clés

Les mots-clés de différentes catégories peuvent potentiellement être liés les uns aux autres. Par exemple les utilisateurs savent qu'il existe une forte relation entre l'*espèce cochon* et la *maladie peste porcine africaine* alors que celle entre *cochon* et *fièvre aphteuse* est très faible, voire inexistante. L'objectif de ce critère est de pouvoir mettre en évidence les relations qui peuvent exister entre des mots-clé provenant de

différentes catégories et notamment de pouvoir pondérer ces relations (liaisons plus ou moins fortes) pour aider à construire des requêtes.

C3 : Gestion des résultats

Ce critère porte sur l'exploration et l'exploitation des résultats de la requête. Les experts ont besoin d'accéder à la requête et à ses résultats. Ils doivent pouvoir sauvegarder les résultats pertinents, modifier l'ordre d'apparition pour par exemple les trier en fonction d'une catégorie, connaître rapidement si les résultats semblent pertinents, modifier simplement la requête pour comparer des résultats, etc.

C4 : Intégration de connaissances externes

Ce critère porte sur l'intégration de connaissances provenant de sources externes telles que des connaissances d'autres experts ou des résultats issus d'approches de fouille de textes [5]. Ces connaissances supplémentaires doivent pouvoir être suggérées à l'expert afin d'enrichir automatiquement les requêtes.

Ces quatre critères seront mis en évidence dans l'état de l'art détaillé dans la section suivante.

2.2.2 État de l'art

De très nombreux travaux se sont concentrés sur la création de requêtes. Par exemple, dans le domaine de la recherche d'informations, les articles [8, 12, 26, 38, 45, 65, 88] proposent des approches pour créer ou enrichir automatiquement des requêtes en utilisant des connaissances externes : ontologies, dictionnaires, bases de connaissances, etc. Cependant, même si ces approches sont très efficaces, elles ne traitent pas la création visuelle de requêtes.

Construction visuelle de requêtes

Les premiers travaux sur la construction visuelle de requêtes concernent des données relationnelles sur lesquelles l'utilisateur peut graphiquement exprimer des requêtes de type SQL [93]. Aujourd'hui, la plupart des systèmes de gestion de bases de données proposent, par exemple, des interfaces tabulaires pour créer des requêtes simples. Par exemple, Chris et al. [90] proposent Polaris, un système pour l'exploration de bases de données multidimensionnelles. Il est fondé sur des tableaux interactifs qui permettent à l'utilisateur de spécifier ses requêtes et de les exécuter.

Récemment des approches ont été proposées pour permettre la construction visuelle de requêtes sur des données de type graphe [11,77,77]. Par exemple, Visage [77] (Cf. Figure 2.4) est un système interactif de construction de requêtes qui permet à l'utilisateur de représenter sa requête sans code complexe. Il utilise une approche guidée par les données ("data-driven"). Cette fonctionnalité permet à l'utilisateur de spécifier les différents types de requêtes. Par exemple, la Figure 2.4 illustre une requête de recherche d'un acteur qui a joué dans deux films réalisés tous-deux par

les deux mêmes réalisateurs. Les requêtes sont construites visuellement en plaçant les nœuds et les arêtes (C2) (Cf. Figure 2.4.a). Les résultats sont affichés dans une liste (Cf. Figure 2.4.b) à droite du graphe de la requête. Quand l'utilisateur clique sur un nœud de résultats, deux fenêtres de résumés sont ouvertes : une pour les caractéristiques (Cf. Figure 2.4.c) et l'autre pour les attributs (Cf. Figure 2.4.d).

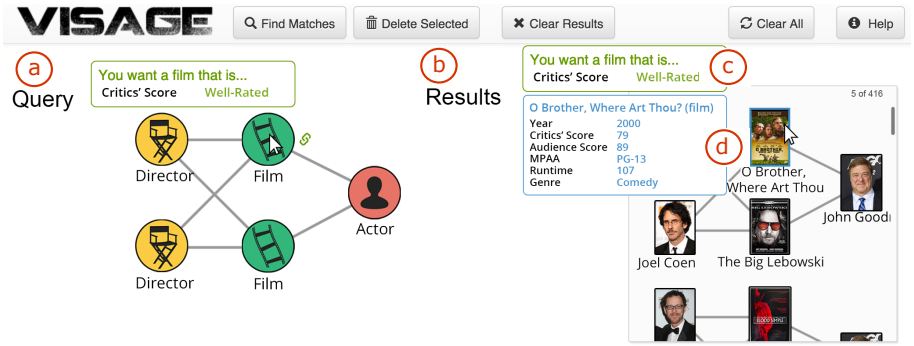


FIGURE 2.4 – Visage : **(a)** expression de la requête, **(b)** résultats de la requête, **(c, d)** attributs d'un nœud sélectionné.

Popoto.js[1] (Cf. Figure 2.5) est une librairie JavaScript qui offre toutes les fonctionnalités pour construire graphiquement des requêtes. Les requêtes correspondent à des graphes interactifs qui peuvent être utilisés pour analyser n'importe quelle application web et les requêtes avancées sont automatiquement créées sur les bases de données Neo4j. La Figure 2.5 illustre une application de Popoto.js dans le cas de données issues de restaurants.

Filtrage des résultats

Orland et al. proposent le système TheHotMap.com [48] (Cf. Figure 2.6), combinant HotMap [50] et Wordbars [49]. Il offre de nombreuses fonctionnalités pour filtrer les résultats des recherches (C3).

L'utilisateur peut par exemple ordonner les résultats en sélectionnant les mots-clés de la requête dans l'arborescence de Worldbars, ou en choisissant les mots-clés fréquents (dans les titres, les extraits ou les URLs) dans les résultats via la liste à gauche de Worldbars. De plus, il peut sélectionner les résultats en utilisant leur aperçu dans la barre de HotMap (les couleurs représentent la fréquence de chaque mot-clé de requête dans chaque résultat). Il est également possible d'ajouter ou de supprimer des mots-clés dans la requête en cliquant sur les boutons "plus" ou "moins" à côté des mots-clés de Wordbars.

1. www.popotojs.com

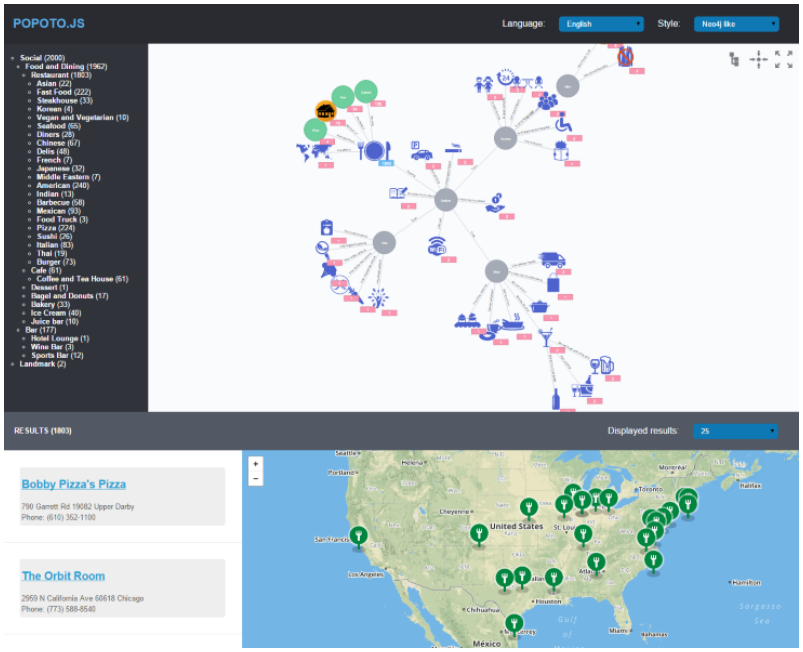


FIGURE 2.5 – Un exemple d'utilisation de la librairie Popoto.js.

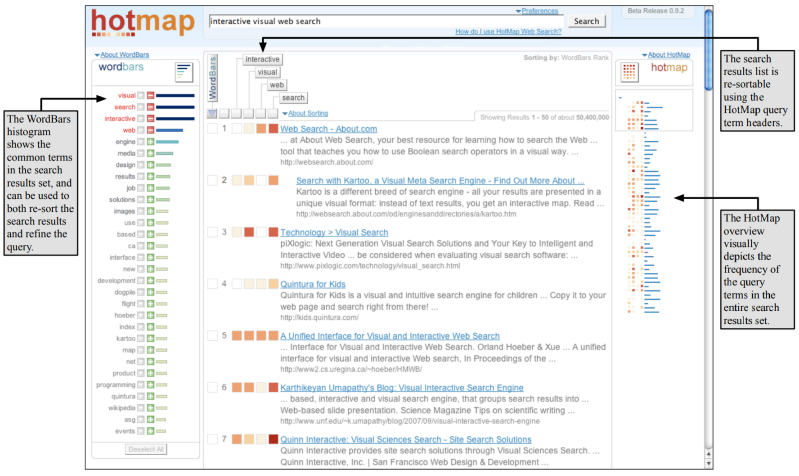


FIGURE 2.6 – Un exemple de filtrage de résultats d'une requête via TheHotMap.com.

Suggestion de nouvelles données

Orland et al. proposent aussi une visualisation de requêtes, VisiQ [51]. L'utilisateur commence par écrire une requête. Ensuite, un graphe est affiché. Il contient les mots-clés de la requête ainsi que d'autres mots-clés et concepts issus d'une base de connaissances (C4). La distance entre les nœuds représente le poids entre les mots-clés et les concepts. L'utilisateur peut sélectionner des mots-clés dans le graphe

et relancer la requête. La Figure 2.7 illustre le graphe de la requête "document clustering" après l'ajout des mots-clés, des concepts et des poids. L'utilisateur peut sélectionner des mots-clés dans le graphe pour construire une nouvelle requête. La requête suggérée dans l'exemple est "clustering information retrieval".

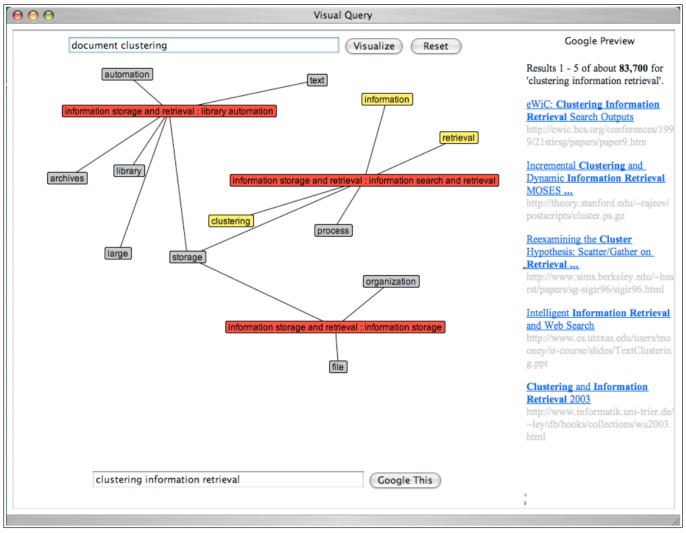


FIGURE 2.7 – VisiQ : le graphe de la requête "document clustering".

Certaines approches proposent également de nouvelles requêtes (C4) à partir de données de type graphe [71, 99] ou non [51, 53, 89]. Par exemple AUTOG [99] offre un mécanisme de suggestion de requêtes de sous-graphes. En fonction de la requête initiale et des données disponibles, le système suggère des affinements que l'utilisateur peut utiliser pour l'enrichir et la relancer.

Discussion

Dans cette section, nous analysons comment les approches présentées préalablement peuvent être utiles pour répondre aux différents critères définis dans la section 2.2.1.

Visage et Popoto.js représentent les mots-clés et leurs relations par des nœuds et des liens dans un graphe. Ils proposent des fonctionnalités proches des critères C1 et C2 pour exprimer les requêtes. Par contre, parmi les fonctionnalités présentées, il n'existe pas de possibilité de définir des poids entre les mots-clés et cette fonctionnalité est importante pour les épidémiologistes.

TheHotMap.com met en évidence le fait qu'il est important d'offrir des résultats avec des indicateurs de pertinence (C3). Ce type d'approche est évidemment très utile et nécessaire pour aider les épidémiologistes à faire face à la quantité de résultats obtenus. Outre ces aspects, il est également indispensable d'offrir des outils

de gestion des résultats (e.g. sauvegarde totale ou partielle des résultats). Enfin, VisiQ, en proposant de nouveaux mots-clés et concepts, permet à l'utilisateur d'intégrer des éléments suggérés (C4). Une approche similaire adaptée au domaine de la veille épidémiologique est pertinente. Dans notre cas, les suggestions ne viennent pas nécessairement d'un dictionnaire mais de sources diverses (autre expert ou procédé de fouille de textes) et surtout peuvent contenir des liens pondérés entre les mots-clés [5].

Dans la section suivante nous proposons EpidVis, notre contribution pour consolider le processus de veille en respectant tous les critères préalablement définis. Cette approche tient compte de certaines propositions issues de l'état de l'art tout en les étendant ou les adaptant à notre contexte.

2.3 EpidVis

EpidVis[2] est une plateforme de création visuelle de requêtes pour la veille en épidémiologie animale. Cette dernière, conçue de manière étroite avec les experts du domaine, propose de nombreuses fonctionnalités pour permettre aux utilisateurs :

1. d'exprimer leurs connaissances en utilisant des mots-clés et leurs relations,
2. de créer facilement et de lancer des requêtes sur le web,
3. d'explorer les résultats,
4. d'intégrer des suggestions à l'aide de connaissances complémentaires.

Ces fonctionnalités sont réalisées via différentes vues (Cf. Figure 2.8). Le gestionnaire de mots-clés (Cf. Figure 2.8.a) contient les mots-clés (C1) et leurs relations (C2) pour permettre aux épidémiologistes d'exprimer leurs connaissances. Le constructeur de requêtes (Cf. Figure 2.8.b) aide à construire et lancer des requêtes basées sur les mots-clés et leurs relations. Les résultats sont affichés dans la visualisation des résultats (Cf. Figure 2.8.c) qui permet de les sélectionner, les filtrer et les enregistrer (C3). Enfin, la visualisation de suggestions (Cf. Figure 2.8.d) propose un enrichissement de l'ensemble des données du gestionnaire de mots-clés en intégrant des connaissances externes (C4). Chacune de ces fonctionnalités est détaillée dans les sous sections suivantes.

2. https://youtu.be/JttBNn-TFj0

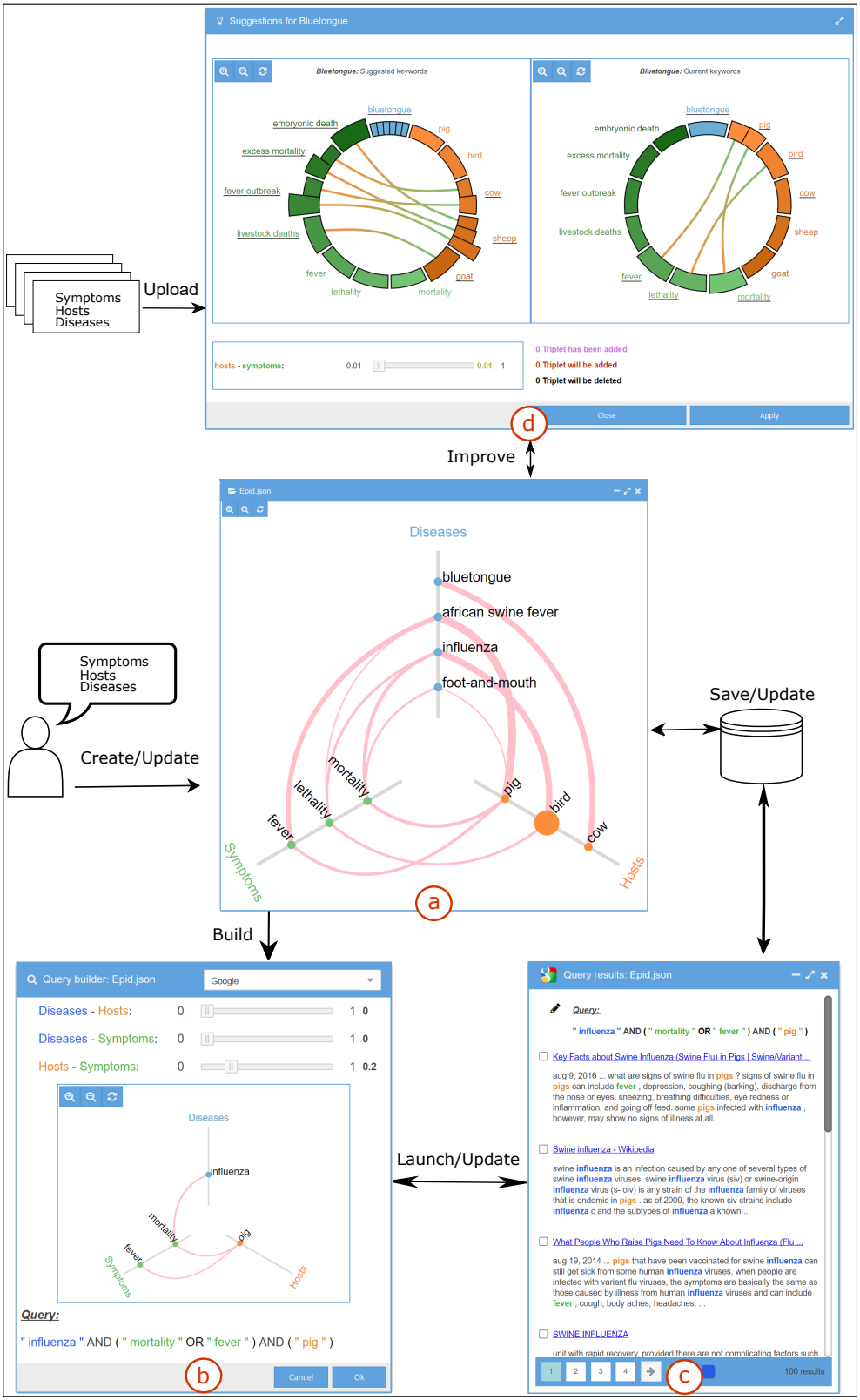


FIGURE 2.8 – EPIDVIS : **(a)** le gestionnaire de mots-clés, **(b)** le constructeur de requêtes, **(c)** la visualisation des résultats et **(d)** la visualisation de suggestions.

2.3.1 Gestionnaire de mots-clés

Le gestionnaire de mots-clés (Cf. Figure 2.9.a) est une vue interactive qui permet aux experts d'exprimer leurs connaissances en utilisant les mots-clés (C1) et leurs relations (C2). Trois catégories sont considérées : les maladies, les symptômes et les espèces. Nous utilisons une visualisation de type *Hive plot* [60] avec trois axes pour représenter ces catégories et leurs mots-clés.

Chaque axe représente une catégorie. Les mots-clés sont représentés comme des nœuds alignés le long des axes correspondants à leur catégorie (C1). Les couleurs des nœuds montrent les catégories, à savoir le bleu pour les maladies, l'orange pour les espèces et le vert pour les symptômes. Des liens de différentes épaisseurs entre les nœuds correspondent aux poids des relations entre les mots-clés (C2). Les nœuds ajoutés apparaissent initialement avec une taille fixe. Comme il est possible de fusionner des mots-clés qui sont sémantiquement similaires (par exemple le mot-clé "*bird*" dans la Figure 2.9), la taille des nœuds fusionnés est proportionnelle au nombre de mots-clés qu'ils contiennent.

FIGURE 2.9 – Gestionnaire de mots-clés et barres d'outils associés. **(a)** Gestionnaire de mots-clés. **(b)** Barre d'outils pour interagir avec le gestionnaire de mots-clés. **(c)(d)** Boutons pour lancer le constructeur de requêtes et la visualisation de suggestions. **(e)** Barre d'outils pour la gestion de fichiers.

De manière à utiliser tout l'espace disponible sur chaque axe, les nœuds sont positionnés uniformément le long de l'axe de la manière suivante. Soit n le nombre

de nœuds, r_i le rayon de chaque nœud i et l la longueur de l'axe. La longueur cumulée des espaces entre les nœuds est $l_{gs} = l - 2 \times \sum_{i=1}^{n} r_i$. On peut donc calculer la longueur de chaque espace entre deux sommets consécutifs de la manière suivante : $l_g = l_{gs}/(n+1)$. Ensuite, la position d'un nœud i le long de son axe est donnée par la fonction :

$$pos_i = i \times l_g + r_i + 2 \times \sum_{j=1}^{i-1} r_j$$

l est dynamique : si $l_{gs} < 0$, *i.e.* l est trop court pour contenir tous les nœuds, alors l est automatiquement allongé pour obtenir $l_{gs} \geq 0$. Lorsque plusieurs nœuds sont fusionnés, nous positionnons le nœud résultant au barycentre de l'ensemble des nœuds fusionnés. Ensuite, nous utilisons une autre fonction [35] pour supprimer les chevauchements (voir chapitre suivant).

Une barre d'outils est proposée (Cf. Figure 2.9.b) pour manipuler l'ensemble des mots-clés et leurs relations. Les utilisateurs peuvent ajouter, supprimer, mettre à jour des relations/mots-clés et fusionner/séparer des nœuds dans chaque catégorie. S'il n'y a pas assez d'espace pour ajouter les nouveaux nœuds, les axes sont automatiquement étendus. Les nœuds peuvent également être déplacés le long de chaque axe et il est possible de visualiser la composition des nœuds fusionnés à l'aide d'un arbre (Cf. Figure 2.10).

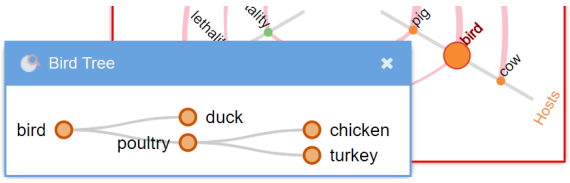


FIGURE 2.10 – Arborescence d'un nœud fusionné : *"bird"* est composé de *"duck"* et de *"poultry"*. *"poultry"* est lui-même composé de *"chicken"* et de *"turkey"*.

Différentes interactions permettent de faciliter l'utilisation. Par exemple la molette de la souris ou les boutons en haut de la vue (zoom avant/arrière/réinitialisation) offrent la possibilité de faire des zooms avant/arrière. Le survol d'un lien affiche une info-bulle indiquant son poids. Un double clic sur les nœuds/liens permet de les mettre à jour les informations (noms, poids).

Enfin, nous pouvons créer, enregistrer ou ouvrir les gestionnaires de mots-clés (Cf. Figure 2.9.e).

2.3.2 Constructeur de requêtes

Le but principal du constructeur de requêtes est d'aider les épidémiologistes à facilement construire et lancer des requêtes sur différents moteurs de recherche. Ces dernières sont générées à partir des mots-clés et des relations entre eux. Initialement,

il faut sélectionner un mot-clé dans le gestionnaire de mots-clés. Ensuite, la vue du constructeur de requêtes est lancée en cliquant sur le bouton dédié (Cf. Figure 2.9.c). La Figure 2.11 montre un exemple de cette vue. Elle contient une proposition de requête contenant le mot-clé sélectionné et des mots-clés supplémentaires (Cf. Figure 2.11.c). Nous expliquons maintenant comment cette requête est construite et comment l'utilisateur peut interagir avec la vue pour l'ajuster.

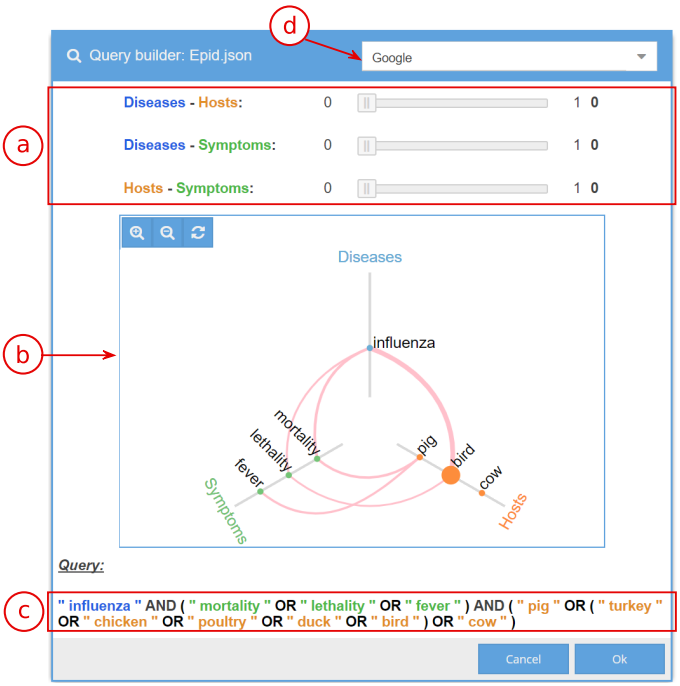


FIGURE 2.11 – Constructeur de requêtes : **(a)** curseurs permettant de filtrer les mots-clés en fonction de la valeur de leurs liens, **(b)** visualisation de la composition de la requête selon les mots-clés et les relations sélectionnés, **(c)** requête générée qui sera lancée et **(d)** moteur de recherche sélectionné.

Les mots-clés utilisés pour construire la requête contiennent celui qui a été initialement sélectionné ainsi que l'ensemble des mots-clés des deux autres axes. Par exemple, si nous avons sélectionné *influenza* dans la Figure 2.9, la requête sera construite avec le sous-ensemble de mots-clés montré dans la Figure 2.11.b. Ces mots-clés sont connectés avec les opérateurs logiques OR et AND. OR joint les mots-clés de la même catégorie, AND joint les mots-clés de différentes catégories (Cf. Figure 2.12). Nous obtenons ainsi la requête représentée dans la Figure 2.11.c.

Nous avons maintenant besoin d'un formalisme pour décrire la requête proposée et comment l'utilisateur peut la modifier. Dans ce qui suit, Ca_1, Ca_2, et Ca_3 sont les trois ensembles de mots-clés correspondants aux trois catégories (maladie, espèce et symptôme). x est le nœud sélectionné, $x \in Ca_1$, et $w(u, v)$ est une fonction qui donne le poids entre les nœuds u and v. Enfin, $Th_{Ca_i Ca_j}$ dénote le seuil pour les liens

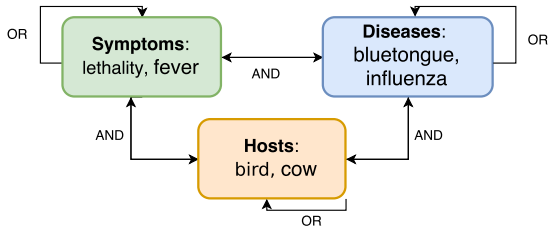


FIGURE 2.12 – Opérateurs logiques utilisés entre les mots-clés des requêtes.

reliant des mots-clés issus des catégories Ca_i et Ca_j (nous verrons ultérieurement comment l'utilisateur peut définir ces seuils).

La requête Q est générée ainsi :

$$
\begin{aligned}
Y_1 &= \{y | y \in Ca_2 \wedge w(xy) \geq Th_{Ca_1Ca_2}\} \\
Z_1 &= \{z | z \in Ca_3 \wedge w(xz) \geq Th_{Ca_1Ca_3}\} \\
Y_2 &= \{y | y \in Y_1 \wedge \exists z \in Z_1 | w(yz) \geq Th_{Ca_2Ca_3}\} \\
Z_2 &= \{z | z \in Z_1 \wedge \exists y \in Y_1 | w(yz) \geq Th_{Ca_2Ca_3}\} \\
Q &= x \cup Y_2 \cup Z_2
\end{aligned}
$$

S'il n'y a pas de liens entre deux mots-clés u et v alors le poids de la relation $(w(x, y))$ vaut 0. Par exemple, la Figure 2.11 illustre une requête où le nœud sélectionné x est la maladie *influenza*, où $Th_{Ca_1Ca_2} = Th_{Ca_1Ca_3} = Th_{Ca_2Ca_3} = 0$ et $Ca_1 =$ maladie, $Ca_2 =$ espèce et $Ca_3 =$ symptôme. Q est représenté dans la Figure 2.11.c. Le mot-clé *"bird"* est un nœud fusionné. Il contient les nœuds : *"turkey"*, *"chicken"*, *"poultry"* et *"duck"*. Ces derniers apparaissent dans la requête Q et sont connectés avec l'opérateur logique OR.

Le constructeur de requêtes contient différents composants. Premièrement, trois curseurs (Cf. Figure 2.11.a) permettent de sélectionner les seuils $Th_{Ca_iCa_j}$. Leur valeur peut varier de 0 à 1. Les mots-clés et leurs relations sont sélectionnés en fonction des seuils comme nous l'avons décrit ci-dessus. Une visualisation des mots-clés sélectionnés et de leurs relations est proposée (Cf. Figure 2.11.b). Nous conservons la même visualisation que dans le gestionnaire de mots-clés. Enfin, la requête est affichée avec les mots-clés liés par les opérateurs logiques (Cf. Figure 2.11.c). Chaque mot-clé a une couleur en fonction de sa catégorie.

Les composants du constructeur de requêtes sont synchronisés. Chaque modification des seuils des relations entre les mots-clés en déplaçant les curseurs met à jour les nœuds/liens de la visualisation de la requête. Par exemple, en partant de la Figure 2.11 et en changeant les seuils (Cf. Figure 2.13.a), les composants du constructeur de requêtes sont mis à jour (Cf. Figure 2.13.b et 2.13.c) selon les règles expliquées précédemment. Dans l'exemple de la Figure 2.13, le seuil entre les maladies et les espèces est de 0,5, celui entre les maladies et les symptômes est de 0 (il

ne doit pas y avoir de lien avec l'axe symptômes), et celui entre les espèces et les symptômes est de 0,1. Cette fonctionnalité permet à l'expert de filtrer les mots-clés en fonction des relations fortes et faibles et met automatiquement à jour la figure mais également la requête.

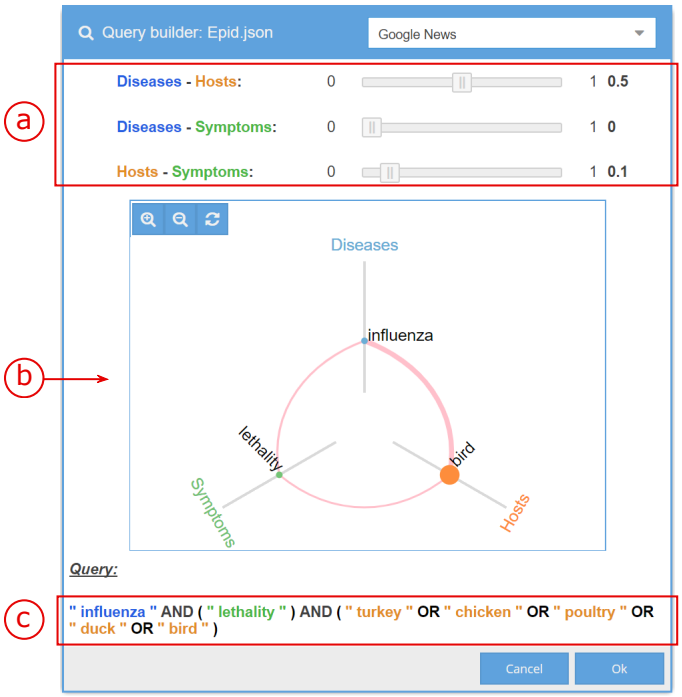


FIGURE 2.13 – Synchronisation des composants du constructeur de requêtes. Le déplacement des curseurs **(a)** modifie la visualisation de la requête **(b)** et la requête **(c)**.

Lorsque la requête est jugée satisfaisante, trois moteurs de recherche sont disponibles pour l'exécuter : Google, Google News et Google avancée (Cf. Figure 2.11.d). En cliquant sur le bouton "Ok", la visualisation des résultats est alors lancée en fonction du moteur de recherche sélectionné. L'utilisation de ces moteurs de recherche est évaluée en section 2.4.3.

2.3.3 Visualisation des résultats

La visualisation des résultats permet aux experts d'analyser les résultats (C3) de la requête qui a été lancée via le constructeur de requêtes. La requête est affichée en haut de la vue (Cf. Figure 2.14.a) avec le même format que celui utilisé dans le constructeur de requêtes. Pour représenter les résultats (Cf. Figure 2.14.b), nous utilisons une liste comme dans les moteurs traditionnels et l'approche mise en place par [51].

FIGURE 2.14 – Visualisation des résultats. **(a)** Requête. **(b)** Résultats sous forme d'une liste. **(c)** Action d'édition qui permet de revenir au constructeur de requêtes. **(d, e)** Boutons de modification de l'ordre des résultats et de changement de page.

Chaque résultat contient un titre et un extrait. Les mots-clés de la requête apparaissent dans les extraits avec les mêmes couleurs que celles des catégories du gestionnaire de mots-clés [48] (Cf. Figure 2.14.a). Par exemple, le deuxième extrait du résultat de la Figure 2.14.b contient les mots-clés *"influenza"* (une maladie), *"cow"* (une espèce), et *"fever"* (un symptôme). Cette représentation permet de savoir rapidement combien de catégories apparaissent dans le résumé et ainsi faire une première sélection des résultats.

Nous associons des cases à cocher à chaque résultat. Il est ainsi possible de pouvoir choisir et enregistrer une sous-liste des résultats (voir aussi les boutons "sauvegarder" et "ouvrir" en haut de la Figure 2.9.e). De plus, des boutons classiques sont disponibles pour changer passer d'une page de résultats à une autre (voir Figure 2.14.e).

Les résultats peuvent être triés via les boutons colorés comme l'illustre la Figure 2.14.d. Ces couleurs se réfèrent aux catégories. Les résultats contenant des extraits ayant des mots-clés des catégories sélectionnées sont positionnés en haut et les résultats ayant des extraits ne contenant aucun mot-clé des catégories sélection-

nées apparaissent à la fin de la liste. Par exemple, la Figure 2.14.a représente la liste ordonnée en fonction de la catégorie maladie (bouton bleu).

Il est possible de modifier la requête en cliquant sur le bouton d'édition (Cf. Figure 2.14.c). Dans ce cas, la vue du constructeur de requêtes s'ouvre avec les paramètres précédents et la requête peut alors être modifiée et ré-exécutée. La Figure 2.15 montre un exemple où les seuils de la requête initiale ont été modifiés. Cette fonctionnalité aide les utilisateurs à raffiner itérativement leurs requêtes.

2.3.4 Visualisation de suggestions

Cette section décrit l'enrichissement des mots-clés en utilisant des connaissances externes (C4). Ces connaissances sont fournies sous la forme de fichiers spécifiques contenant des mots-clés ainsi que de leurs relations, e.g. les interactions entre des espèces et des symptômes associés à une maladie spécifique.

2.3.4.1 La représentation visuelle

La visualisation de suggestions nécessite dans un premier temps de sélectionner, dans le gestionnaire de mots-clés, un mot de n'importe quelle catégorie (Cf. Figure 2.9.d). La vue apparaît après avoir sélectionné un fichier contenant des informations relatives à ce mot-clé. La Table 2.1 illustre un exemple de fichier contenant des informations associées à la maladie *bluetongue*. Le poids, colonne Weigh, représente le poids du lien entre un symptôme et une espèce. Il est directement renseigné par un expert ou obtenu à l'aide d'un processus semi-automatique combinant des approches de fouille de textes et fouille de web [5].

Symptom	Host	Weight
embryonic death	sheep	30
excess mortality	cow	60
.....	.....	....

TABLE 2.1 – Exemple de fichier contenant des symptômes et des espèces associés à la maladie *bluetongue*.

La visualisation de suggestions comporte deux vues principales : 1) la vue des mots-clés suggérés à gauche (Cf. Figure 2.16.a) contient les mots-clés et les relations contenus dans le fichier externe, 2) la vue de mots-clés disponibles à droite (Cf. Figure 2.16.b) contient les mots-clés et les relations déjà présents dans le gestionnaire de mots-clés.

Les données des deux vues sont représentées sous la forme d'un cercle divisé en arcs. Ces arcs représentent les mots-clés. Afin de rendre les cercles homogènes, les mots affichés correspondent à l'union des mots-clés disponibles dans le gestionnaire de mots-clés et des mots-clés suggérés. De manière à mettre en évidence les mots

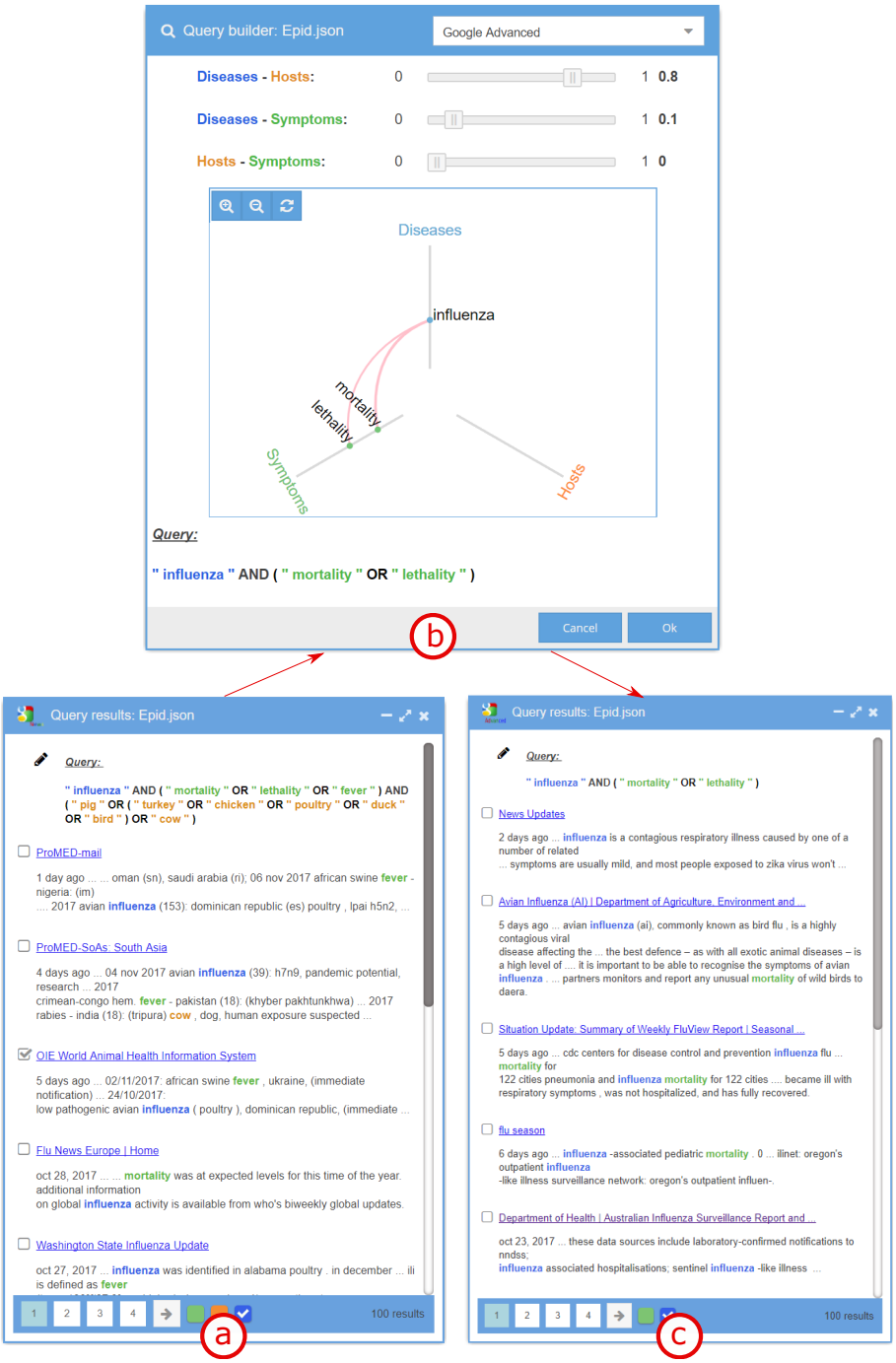


FIGURE 2.15 – Raffinement itératif de la requête. **(a)** Visualisation des résultats d'une requête initiale. **(b)** Modification de la requête en utilisant les curseurs ou en modifiant le moteur de recherche. **(c)** Nouveaux résultats obtenus à partir de la requête mise à jour.

FIGURE 2.16 – Visualisation de suggestions. **(a)** Relations et mots-clés suggérés à partir d'un fichier externe. **(b)** Relations et mots-clés disponibles dans le gestionnaire de mots-clés. **(c)** Curseur permettant de filtrer les relations et les mots-clés en fonction de leur poids. **(d)** Résultats des différentes actions réalisées.

disponibles, ces derniers sont soulignés (Cf. Figure 2.16). Le mot-clé sélectionné apparaît au sommet des cercles (arc bleu "bluetongue" dans l'exemple) et les autres mots-clés sont visualisés sur le pourtour des cercles. Chaque arc possède un nom et une couleur en fonction de sa catégorie. Un dégradé de couleurs est utilisé pour différencier les mots-clés de la même catégorie. Les relations entre les mots-clés sont représentées à l'aide de lignes entre les arcs correspondants. De manière à ne pas surcharger la visualisation, le poids des liens est reporté sur les arcs associés aux mots-clés. Chaque lien a une interpolation de couleur inversée entre les catégories des arcs reliés. Enfin, afin d'éviter les encombrements sur le dessin, les liens entre le mot-clé sélectionné et les autres mots-clés sont initialement cachés et n'apparaissent qu'au survol de la souris ou bien quand ils sont sélectionnés.

Un curseur (Cf. Figure 2.16.c) peut être utilisé pour filtrer les mots-clés et les relations suggérés en fonction du poids des relations. Comme de nombreuses opérations peuvent être effectuées avant d'être reportées sur le gestionnaire de mots-clés, une information supplémentaire décrite les actions réalisées lors des manipulations (Cf. Figure 2.16.d). Chaque action est codée avec une couleur spécifique : la couleur violette pour les triplets (maladie, espèce, symptôme) déjà ajoutés, la couleur rouge pour les triplets sélectionnés à ajouter au gestionnaire de mots-clés et la couleur noire pour les triplets à supprimer.

2.3.4.2 Les interactions de la visualisation de suggestions

De nombreuses interactions sont proposées pour explorer les données suggérées et les ajouter au gestionnaire de mots-clés. Le survol d'un mot-clé met en évidence le mot, ses voisins et les relations qu'ils partagent (Cf. Figure 2.17.a). Dans ce cas, les arcs sont divisés et les relations entre le mot survolé et les autres mots-clés sont affichées. Ces derniers forment toujours un triplet maladie/espèce/symptôme.

Les deux vues sont synchronisées : le survol d'un triplet dans la vue de mots-clés suggérés montre ce même triplet dans la vue de mots-clés disponibles. Par exemple, dans la Figure 2.17.a, le triplet survolé dans la vue de mots-clés suggérés; *"bluetongue"/"sheep"/"fever outbreak"*, est également affiché dans la vue de mots-clés disponibles (Cf. Figure 2.17.b). Lorsque la souris ne survole plus le mot-clé, le triplet est supprimé de la vue de mots-clés disponibles et la visualisation de suggestions revient à son état initial.

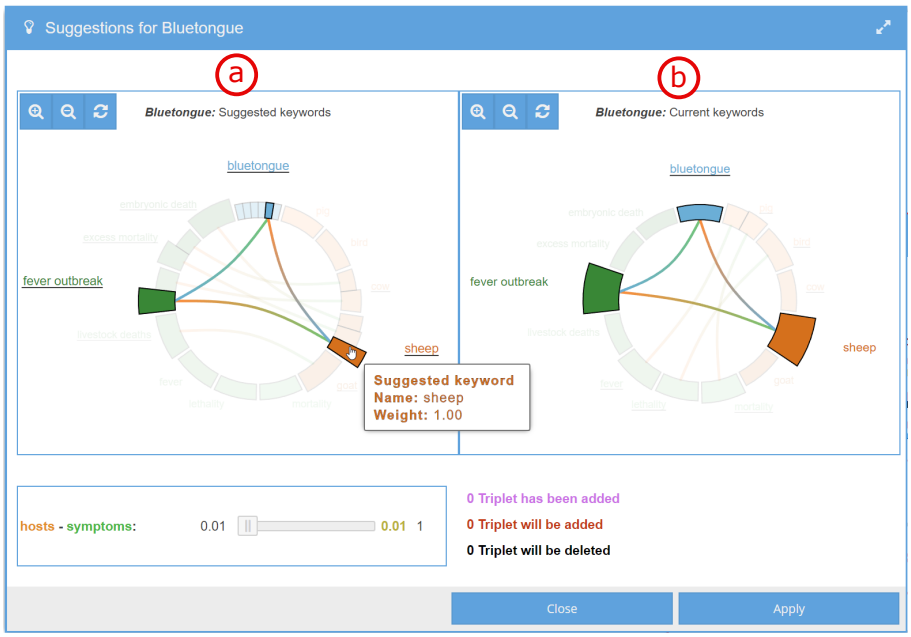


FIGURE 2.17 – Survol d'un mot-clé dans la vue de mots-clés suggérés.

Il est possible d'ajouter ou de supprimer des triplets dans le gestionnaire de mots-clés en utilisant la visualisation de suggestions. Toutes les vues sont synchronisées : la vue de mots-clés suggérés, la vue de mots-clés disponibles et le gestionnaire de mots-clés. Ainsi, les utilisateurs peuvent sélectionner des triplets dans la vue de mots-clés suggérés en cliquant sur les arcs (Cf. Figure 2.18). Ces triplets sélectionnés sont représentés avec une couleur rouge et apparaissent dans les deux vues de suggestions. Ils peuvent être ajoutés dans le gestionnaire de mots-clés en utilisant le bouton "Apply" (Cf. Figure 2.18.a). Dans ce cas, la vue de mots-clés disponibles et le gestionnaire de mots-clés sont mis à jour avec les nouveaux triplets. Les liens

des nouveaux triplets ajoutés dans le gestionnaire de mots-clés apparaissent avec la couleur violette. À partir de la vue suggestion, il est possible de dé-sélectionner un triplet ajouté qui apparaît alors en noir (Cf. Figure 2.18.a). Ces triplets seront alors supprimés automatiquement lors du prochain "Apply" (Cf. Figure 2.18.b). L'information est automatiquement mise à jour en fonction des actions : sélection, ajout et suppression (Cf. Figure 2.18.a, 2.18.b).

FIGURE 2.18 – Interactions avec le gestionnaire de mots-clés : **(a)** les triplets déjà ajoutés au gestionnaire de mots-clés apparaissent en violet, les triplets à supprimer apparaissent en noir et les triples à ajouter en rouge, **(b)** mises à jour après l'action de "Apply".

Les deux vues sont également synchronisées pour les actions de zoom avant/arrière/réinitialisation via les boutons supérieurs de chaque vue, ou les changements de zoom à l'aide de la molette de la souris.

Le curseur du bas de la visualisation permet de filtrer les mots-clés en fonction du poids des relations. Les liens et les mots-clés associés ayant une valeur inférieure à celle choisie sont supprimés de la visualisation. La Figure 2.19 illustre les mises à jour des visualisations lorsque le seuil sélectionné est de 0.22.

FIGURE 2.19 – Mises à jour des vues lors d'un filtrage de données par la sélection d'un seuil sur le poids des relations (passage de 0,01 à 0,22).

2.3.5 Considérations techniques

EpidVis est une application web. L'interface d'utilisateur a été implémentée en Javascript à l'aide de la bibliothèque Extjs [3]. Les différentes vues ont été créées en utilisant la bibliothèque D3.js [14]. Un serveur PHP est utilisé pour exécuter les requêtes et obtenir les résultats. Enfin, la connexion entre l'interface de l'utilisateur et le serveur est réalisée à l'aide de la bibliothèque JQuery [4].

2.4 Évaluation

Dans cette section, nous évaluons EpidVis selon trois axes complémentaires. Dans un premier temps, nous analysons comment les utilisateurs appréhendent l'outil. Nous présentons ensuite une étude de cas menée par une experte à fin d'illustrer comment EpidVis permet d'accompagner le travail de veille. Cette étude met également en évidence le comportement des moteurs de recherche en fonction des mots-clés ajoutés et surtout des combinaisons de ces derniers avec les opérateurs logiques. Enfin, nous proposons quelques expérimentations pour étudier le comportement des différents moteurs de recherche en fonction de la complexité des requêtes.

2.4.1 Étude utilisateurs

De manière à évaluer l'utilité et la facilité d'utilisation d'EpidVis, un questionnaire [5] a été soumis à un groupe de 12 participants. Pour chacune des vues, différentes questions ont été posées comme par exemple :

— En utilisant le gestionnaire de mots-clés, avez-vous trouvé la fusion et la séparation des nœuds utiles/faciles à utiliser ?

— En utilisant le gestionnaire de requêtes, avez-vous trouvé la modification de requêtes utile/facile à utiliser ?

— En utilisant la visualisation de suggestions, avez-vous trouvé la synchronisation entre les deux vues utile/facile à utiliser ?

Le but de cette étude n'était pas de montrer comment l'outil aide les experts du domaine dans leurs tâches quotidiennes (voir la section suivante pour des exemples), mais d'évaluer comment les différentes vues facilitaient l'accomplissement de tâches générales. Par exemple, dans la visualisation de suggestions, le but était d'évaluer la valeur ajoutée pour i) intégrer des données supplémentaires issues d'un fichier externe et ajouter des données intéressantes manuellement à l'outil (utilité), et ii) utiliser des fonctionnalités interactives pour intégrer de nouvelles données dans le gestionnaire de mots-clés (facilité d'utilisation).

3. www.sencha.com/products/extjs
4. http://jquery.com/
5. https://goo.gl/forms/E7stx5TOh2xfSePf2

Comme aucune compétence particulière en épidémiologie n'était requise pour les 12 participants, il y avait trois experts en épidémiologie et neuf personnes non expertes. D'abord, nous leur avons fait une démonstration détaillée de l'outil et des différentes fonctionnalités. Ensuite, ils ont été invités à se familiariser avec l'outil. Il n'y avait pas de contrainte de temps pour cette tâche, car ils avaient un accès en ligne à l'outil pendant plusieurs jours. Lorsqu'ils étaient prêts, ils devaient remplir un questionnaire en ligne sur la facilité d'utilisation et l'utilité de l'outil. Par exemple, les participants ont été invités à manipuler la visualisation de suggestions : ouvrir un fichier externe, ajouter / supprimer des mots-clés et des relations, etc. Ensuite, il leur a été demandé de noter l'utilité et la facilité d'utilisation de chaque fonctionnalité sur une échelle de 1 à 5 (1 étant "non utile" / "difficile à utiliser", 5 étant "très utile" / "très facile à utiliser"). Nous avons également recueilli des commentaires informels pour chaque vue.

La Table 2.2 montre les résultats moyens de l'évaluation. Aucune différence significative n'est apparue entre les experts du domaine et les non experts, ils ne sont donc pas séparés dans le tableau. Les participants ont apprécié les différentes visualisations en termes d'utilité (note moyenne : 4,2/5) et de facilité d'utilisation (note moyenne : 4,3/5). En particulier, l'utilité du gestionnaire de mots-clés a obtenu les notes les plus élevées (4,5/5 en moyenne). Comme prévu, la vue des résultats de la requête, qui liste principalement les résultats (tout comme les moteurs de recherche classiques) a obtenu la note d'utilité la plus basse (3,6/5 alors que les autres ont des scores entre 4,2 et 4,5).

Les vues d'EpidVis	**Utilité (/5)**	**Facilité d'utilisation (/5)**
Le gestionnaire de mots-clés	4.5	4.3
Le constructeur de requêtes	4.4	4.3
La visualisation des résultats	3.6	4.5
La visualisation de suggestions	4.2	4.1

TABLE 2.2 – Evaluation qualitative d'EPIDVIS : scores moyens de tous les participants.

Les commentaires laissés par les participants soulignent que les fonctionnalités de filtrage et d'enregistrement des résultats ont été particulièrement appréciées. Ils n'ont pas mis en évidence d'autres besoins pour l'outil ni de limitation spécifique en ce qui concerne son utilité ou sa facilité d'utilisation, si ce n'est l'inclusion d'autres moteurs de recherche comme Bing ou Yahoo. Notre système peut gérer ces moteurs mais nous n'avons pas eu un accès gratuit à leurs API pendant le développement du prototype.

2.4.2 Étude de cas

L'objectif de l'étude de cas est de voir si EpidVis offre une réelle aide aux épidémiologistes dans leur travail de veille. Elle a été réalisée par une experte [6] en intelligence épidémiologique pour la plateforme française de veille de la santé animale [7].

Le domaine d'expertise retenu a été la *peste porcine africaine*. Trois tâches ont été définies.

— **Tâche 1 :** l'objectif est de permettre à l'experte de s'approprier les fonctionnalités d'EpidVis et de faire une première requête qui servira de base de comparaison dans les tâches suivantes. Pour cela la requête porte uniquement sur le nom de la maladie étudiée (*peste porcine africaine*).

— **Tâche 2 :** l'objectif est de s'intéresser à l'ajout de mots-clés sur une catégorie. Dans ce cadre un intérêt particulier a été porté sur la possibilité de pouvoir prendre en compte plusieurs langues. En effet, dans leur recherche de dépêches, les experts doivent régulièrement utiliser différentes langues. La catégorie retenue dans cette tâche est les maladies dans laquelle l'experte a décliné le nom de la maladie étudiée dans plusieurs langues.

— **Tâche 3 :** l'objectif est de combiner automatiquement différents mots-clés à l'aide d'opérateurs AND (plusieurs catégories) et OR (plusieurs mots-clés d'une même catégorie) afin d'étudier les résultats obtenus.

Les requêtes des différentes tâches ont été exécutées à l'aide de moteur de recherche Google Actualités. Les dix premières dépêches de chaque requête ont été analysées par l'experte pour évaluer leurs pertinences, i.e. une apparition réelle d'épidémie ou de mesures de contrôle pour la maladie étudiée.

2.4.2.1 Tâche 1 : recherche d'une seule maladie

L'experte a démarré d'un gestionnaire de mots-clés vide et y a ajouté un seul mot-clé, *"peste porcine africaine"*, sur l'axe des maladies. Elle a ensuite lancé la requête à partir de ce mot, ce qui correspond à une requête de base sur un moteur de recherche.

Sur les dix premiers résultats, seuls deux résultats ont été jugés pertinents par l'experte. Parmi les non pertinents, trois d'entre eux étaient des sites web contenant le mot-clé mais ne présentant pas d'intérêt, quatre concernaient des associations d'agriculteurs et un était un site de traduction (Linguee). La moyenne des délais des informations retournées était de 17,9 jours (variation de 2 à 29 jours).

6. Alizé Mercier, ASTRE, Cirad, Inra,Montpellier, France
7. ESA Platform - https://plateforme-esa.fr/

2.4.2.2 Tâche 2 : ajout de différents mots-clés dans une catégorie

Étant donné que l'Allemagne et la Pologne partagent une frontière commune et que la *peste porcine africaine* circule en Pologne depuis février 2014 [84], les langues suivantes ont été retenues : le français (*"peste porcine africaine"*), l'allemand (*"Afrikanische Schweinepest"*) et le polonais (*"afrykańskiego pomoru świń"*). L'experte a donc ajouté ces trois mots-clés sur l'axe des maladies. Puisqu'ils concernent la même maladie, elle les a regroupé à l'aide de la fonctionnalité de fusion de nœuds (Cf. Figure 2.20.b). Elle a choisi de nommer le nœud fusionné (Cf. Figure 2.20.a) avec le nom en français. Elle a ensuite lancé la requête. Comme nous pouvons le constater sur la Figure 2.20.c, les mots ont été automatiquement regroupés avec un OR logique dans la requête.

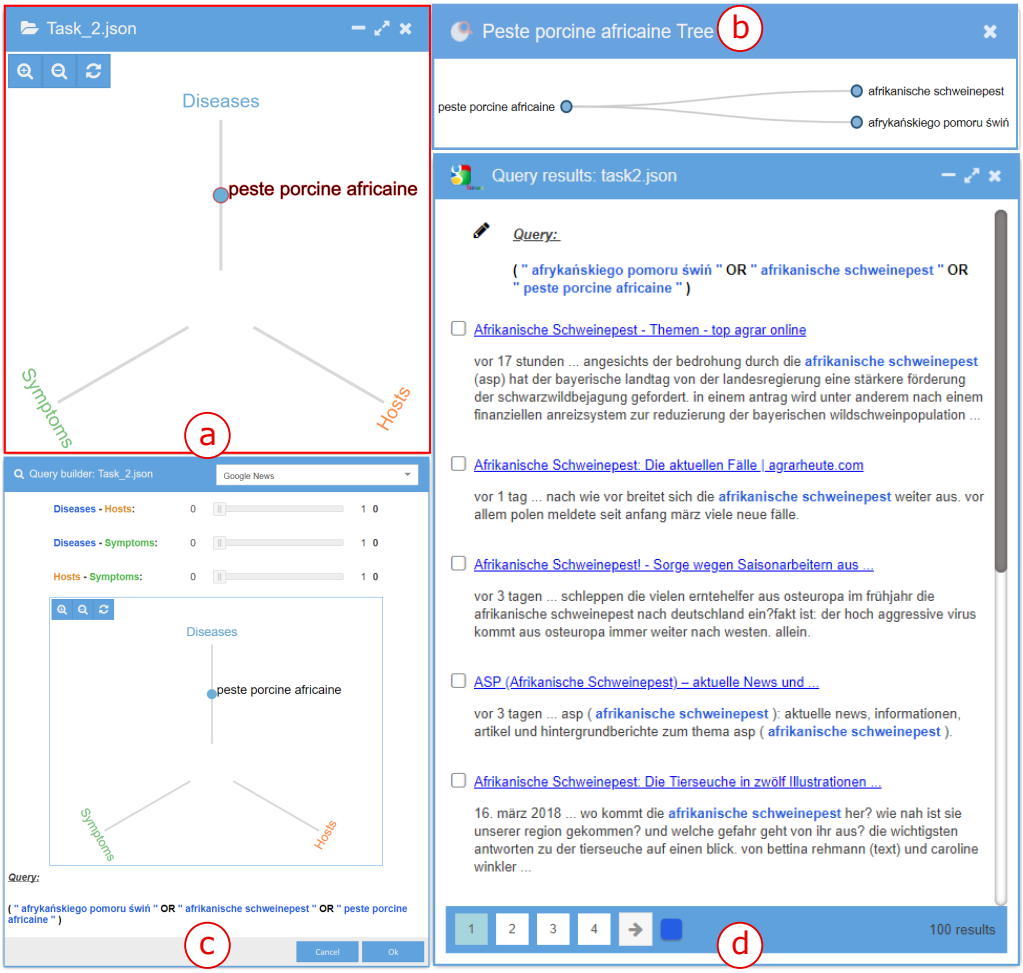


FIGURE 2.20 – Construction de la requête et résultats. La requête est construite à partir de plusieurs mots-clés d'une même catégorie.

Sur les dix premiers résultats (Cf. Figure 2.20.d) huit d'entre eux étaient pertinents. Les deux résultats non pertinents concernaient le blog d'un chasseur et un document présentant la maladie et comment éviter la propagation d'un foyer. Sur les dix résultats, neuf étaient en allemand (dont deux venaient de médias Autrichiens) et un était en polonais. Lors des expérimentations, elle a également voulu intégrer l'anglais dans la recherche. Il a été intéressant de noter que dès que cette langue était prise en compte, la majorité des résultats étaient en anglais reléguant très loin dans la liste les résultats dans d'autres langues. Ceci est principalement lié au fait que de très nombreux documents sont disponibles en anglais.

2.4.2.3 Tâche 3 : ajout de mots-clés dans plusieurs catégories

De manière à prendre en compte les liens entre les catégories, l'experte a enrichi son jeu de données en ajoutant des mots-clés sur les deux autres axes : "*mortalité*", "*hémorragie*" et "*fièvre*" sur l'axe des symptômes, "*cochon*" et "*sanglier*" sur l'axe des espèces. Elle a également ajouté des liens pondérés (Cf. Figure 2.21.a). En sélectionnant dans le gestionnaire de mots-clés la *"peste porcine africaine"*, la requête comportant les opérateurs AND et OR a été générée (Cf. Figure 2.21.b).

Sur les dix premiers résultats (Cf. Figure 2.21.c), quatre d'entre eux ont été jugés pertinents par l'experte. Il est important de noter qu'aucun de ces résultats n'apparaissaient dans la tâche 1. Parmi les 6 résultats non pertinents, deux résultats apparaissaient également comme non pertinents dans la tâche 1 indiquant une redondance entre les deux tâches. Les autres résultats non pertinents étaient des sites d'information générale (Cf. résultats tâche 2) ou une page Wikipédia.

2.4.2.4 Conclusions sur l'étude de cas

À partir de cette analyse, l'experte a tout d'abord conclu qu'EpidVis était facile à utiliser et assez intuitif. La possibilité de pouvoir modifier facilement les requêtes en filtrant les valeurs de liens pour ajouter ou supprimer des mots-clés a également été très apprécié car elle a mis en évidence le fait qu'il était facile et rapide d'avoir des premiers résultats et de pouvoir affiner la requête. La visualisation des liens entre mots-clés et la fusion ont également été notés comme des éléments importants.

L'analyse des résultats retournés a permis de mettre en évidence la valeur ajoutée de la prise en compte des noms des maladies dans différentes langues, en particulier les langues des pays affectés ou menacés. Le nombre de résultats pertinents a été multiplié par quatre. Le fait de prendre en compte les différents axes (tâche 3) a montré qu'il était possible de faire apparaître de nouveaux résultats pertinents (deux fois plus de résultats que pour la première tâche). Le fait de ne pas retrouver le résultat de la tâche 1 dans la tâche 3 pourrait sembler être une limite à l'approche. Cependant nous avons noté que l'experte dans l'utilisation d'EpidVis a tendance à partir de la requête la plus complexe (i.e. après avoir positionné toutes les informations dans le gestionnaire de données), puis d'analyser les résultats, sauvegarder les plus pertinents et continuer en affinant la requête à l'aide des curseurs de filtrage. En

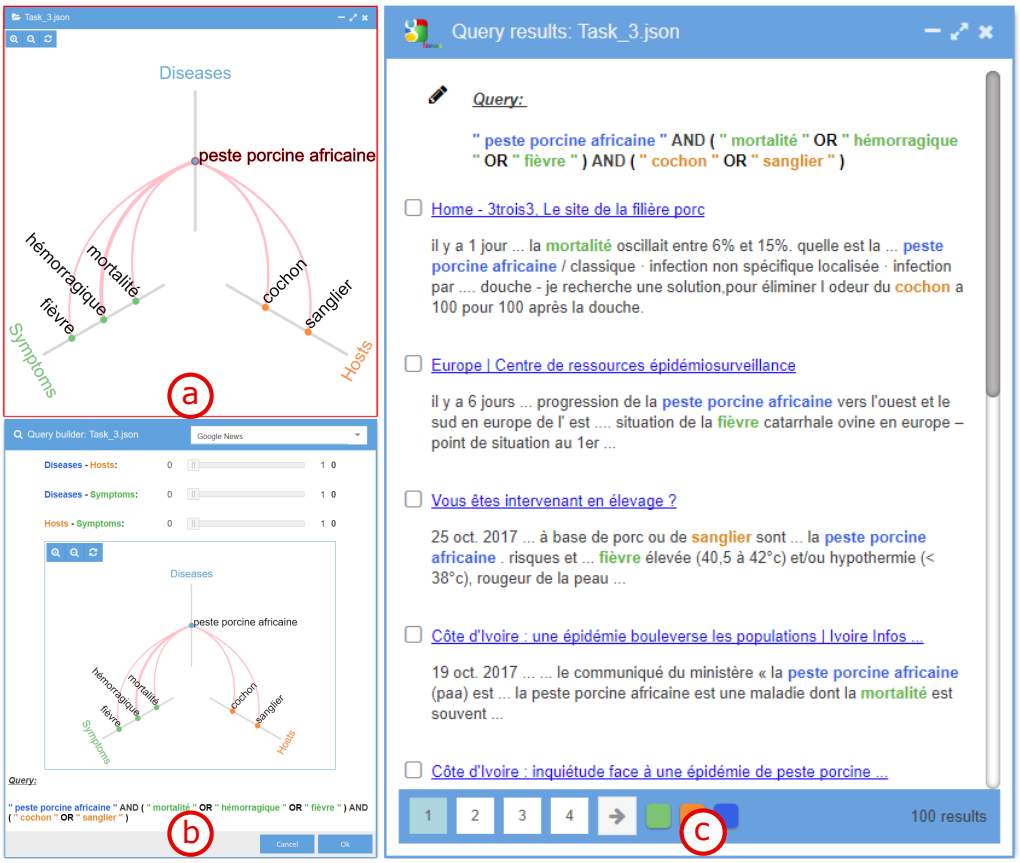


FIGURE 2.21 – Construction de requête et résultats. La requête est construite à partir des mots-clés des trois catégories.

d'autres termes elle finit généralement par des mots-clés issus d'une seule catégorie (tâche 2 puis tâche 1).

2.4.3 Discussion sur les requêtes complexes

Grâce au constructeur de requêtes, les experts peuvent générer des requêtes complexes en ajoutant des mots-clés à l'aide d'opérateurs logiques AND et OR. Lors de nos expérimentations, nous avons pu constater que les différents moteurs de recherche ne les traitent pas de la même manière. Nous avons donc souhaité réaliser des expérimentations supplémentaires afin d'avoir un meilleur aperçu de leur comportement. Pour cela, nous avons lancé des requêtes sur différents moteurs de recherche et comparé leurs résultats. Le but était d'étudier l'impact de l'augmentation de la complexité des requêtes sur la liste des pages web retournées : l'ajout de nouveaux termes entraîne-t-il des changements importants à la liste des résultats ? Cette tâche

a été réalisée par les développeurs d'outils et n'a pas nécessité l'intervention d'experts du domaine.

Une des difficultés pour mener à bien une telle comparaison est d'identifier les différentes URL retournées par les moteurs qui pointent en réalité vers les mêmes sites Internet. De nombreuses approches de la littérature ont été proposées pour comparer les résultats de pages web afin de s'assurer qu'elles ne se réfèrent pas aux mêmes sites. Elles peuvent être classées en deux catégories : 1) celles qui se concentrent sur le contenu ou sur extraits de la page [2] ; 2) celles qui ne considèrent que l'URL [9, 63]. Dans notre cas, nous utilisons le deuxième type d'approche. La méthodologie consiste à 1) normaliser l'URL, 2) générer des règles et identifier les URL similaires en utilisant la détection Dust, et 3) calculer un taux de similarité entre les résultats de la requête. Les requêtes ont été construites en utilisant des combinaisons de mots-clés fréquemment utilisés par les experts appartenant aux catégories suivantes :

— **maladie** : *flu outbreaks, avian flu, influenza* ;

— **espèce** : *poultry, chicken, turkey* ;

— **symptôme** : *lethality, losses, mortality.*

D'abord, une requête est construite en choisissant un mot-clé dans une catégorie (par exemple, *influenza* issu de la catégorie maladie). L'extension de la requête se fait en ajoutant des opérateurs AND (ajout de mots-clés d'autres catégories), ou OR (ajout de mots-clés de la même catégorie) de manière progressive (par exemple, *influenza AND turkey*, *influenza OR avian flu*). Pour chaque moteur de recherche, nous commençons par une requête simple, puis nous complexifions cette requête, puis nous comparons le nombre moyen de résultats identiques entre les deux requêtes (simple et complexifiée), et ainsi de suite. Par exemple, nous comparons la moyenne des résultats similaires pour les deux requêtes : *influenza* et *influenza AND turkey.*

La Table 2.3 présente les résultats [8] en considérant différentes requêtes. Le processus est exécuté comme suit : pour un mot-clé dans une catégorie (e.g. *influenza*), nous étendons la requête en ajoutant de manière itérative un mot-clé d'une autre catégorie (e.g. *influenza* AND *turkey*) ou des deux autres catégories (e.g. *influenza* AND *turkey* AND *mortality*). Parallèlement, pour chaque catégorie, la requête est aussi étendue en ajoutant d'autres mots-clés de la même catégorie avec un opérateur OR (e.g. *influenza* OR *bird flu* AND *turkey* OR *birds* AND *mortality* OR *fever*). Cela nous permet de calculer toutes les combinaisons et de rapporter la moyenne des résultats qui se chevauchent.

Parmi les trois moteurs de recherche les plus utilisés (Google, Bing et Yahoo), Google fournit les meilleurs résultats tout en se basant sur le langage de requête le plus flexible [79]. C'est pourquoi, pour notre évaluation, nous avons utilisé trois moteurs de recherche Google : Google, Google News (se référant uniquement aux actualités) et Google Advanced (qui permet d'utiliser des filtres plus spécifiques).

8. Tous les résultats sont disponibles sur https://goo.gl/BYvdQ4

Nous pouvons observer que l'extension d'une à deux ou trois catégories de mots-clés donne plus de résultats qui se chevauchent plus. Par exemple, le chevauchement des résultats de Google dans une ou deux catégories augmente de 63% à 71%. Les résultats de Google Actualités diminuent légèrement de 78% pour une catégorie à 76% pour deux catégories, mais augmentent à 77% pour trois catégories. Notez que Google Advanced a 100% pour trois catégories, c'est-à-dire que dans ce contexte, les résultats convergent vers les mêmes pages Web. Ainsi, nous recommandons d'utiliser Google Advanced qui obtient une bonne convergence en fonction du nombre de catégories.

Types des requêtes	**Google** Moyenne (Variance)	**Google Actualités** Moyenne (Variance)	**Google Avancé** Moyenne (Variance)
Une catégorie	63% (10%)	78% (9%)	71% (10%)
Deux catégories	71% (10%)	76% (7%)	85% (4%)
Trois catégories	85% (2%)	77% (4%)	100% (0%)

TABLE 2.3 – Résultats d'enrichissement de requêtes pour trois moteurs de recherche en faisant varier le nombre de catégories à considérer dans les requêtes.

En conclusion, tandis que l'évaluation qualitative (section 2.4.2) basée sur l'étude de cas met en évidence l'impact du choix des mots-clés liés aux différentes catégories (maladies, espèces, symptômes), l'évaluation quantitative (section 2.4.3) complète l'évaluation qualitative en mesurant la manière dont le type de combinaisons (catégories de mots-clés et moteurs de recherche) peut influencer la convergence des résultats.

2.5 Conclusion

Dans ce chapitre nous avons présenté l'approche EPIDVIS dont l'objectif est de fournir un outil d'aide à la veille épidémiologiques. Il est important de noter que les différentes vues proposées ont été réalisées conjointement avec les utilisateurs. Le fait d'avoir eu de nombreuses réunions nous a permis, non seulement de mieux comprendre les besoins, mais surtout de pouvoir communiquer rapidement sur les critères qui étaient indispensables. Ces critères ont guidé de nombreux choix de visualisation comme cela a été mis en valeur tout au long de ce chapitre. Les versions présentées dans ce chapitre sont celles qui ont été validées et qui sont actuellement utilisées par les experts.

Concevoir une visualisation adaptée aux besoins des utilisateurs nécessite de nombreuses interactions. Dans la section suivante nous revenons sur les différentes évolutions concernant la représentation retenue pour les suggestions. Une discussion est ensuite proposée sur les différentes techniques d'étiquetage que nous avons envisagées. Enfin nous revenons sur un problème de chevauchement apparu lors de la fusion de plusieurs nœuds dans le gestionnaire des mots-clés.

2.5.1 Évolution de la visualisation de suggestions

Comme nous l'avons vu précédemment, l'un des objectifs de ce processus est, à partir d'un mot-clé, de rechercher des informations extérieures qui lui sont associées afin de pouvoir les suggérer à l'utilisateur. Une première approche a été proposée mettant en évidence le mot-clé au sommet d'un cercle et la répartition des informations des deux autres axes autour de ce cercle. Le premier niveau de l'arborescence dans la Figure 2.22 illustre les premières visualisations de suggestions que nous avons proposé. Nous pouvons constater que les éléments peuvent être répartis facilement et offrir une bonne visualisation générale des différents mots-clés et de leur répartition. Dans ces représentations chaque mot-clé d'un axe peut être rattaché à un autre axe via un lien. Dans ce cas, à chaque occurrence d'un mot-clé du fichier externe correspond un lien.

Dans la première représentation (Figure 2.22.a), chaque mot-clé est représenté par une suite de points placés le long d'un arc de cercle. Le nombre de points correspond au nombre de liens du mot-clé. La taille des points permet de visualiser la pondération du lien. Ainsi, nous n'avons pas besoin d'utiliser l'épaisseur des liens pour montrer cette information et nous ne surchargeons pas le centre du cercle. Dans ce contexte, des problèmes de chevauchement et des difficultés d'interactions ont été observés. En effet pour sélectionner une suggestion, il est indispensable de sélectionner un point et ceci peut être complexe étant donnée leur proximité et leur taille.

Dans la deuxième représentation à droite (Figure 2.22.b), des boites englobantes des points de la première proposition représentent les mots-clés et les poids sont affectés aux liens. L'interaction se fait à présent en sélectionnant les liens plutôt que les mots-clés mais il s'avère qu'elle est difficilement utilisable lorsque le nombre de liens augmente.

Nous avons alors proposé une troisième représentation qui fusionne les deux approches précédentes (Cf. Figure 2.22.c). Elle utilise une boîte englobante contenant des petits cercles pondérés. Dans ce cas, nous évitons le problème de surcharge de liens et de nœuds. Néanmoins, les problèmes de la première proposition persistent.

L'approche finale (Cf. Figure 2.22.d) a consisté à remplacer les points par des arcs dans la boîte englobante. Les poids des relations sont donc représentés par la profondeur de ces arcs. Les liens sortant du mot-clé *bluetongue* sont cachés, ils ne deviennent visibles que pendant les interactions (e.g. survol de la souris) sur les liens.

2.5.2 Étiquetage d'objets placés le long d'un cercle

Plusieurs approches ont été utilisés pour étiqueter des objets placés le long d'un cercle (e.g. [10, 18, 19, 22, 30, 34, 61, 82, 87, 100]). Il existe traditionnellement quatre approches principales. Nous avons développé chacune d'entre elles pour pouvoir, avec les utilisateurs, offrir celle qui était le plus adaptée. Ces techniques ont été présentées aux experts pour en choisir celle qui leur convenait le plus.

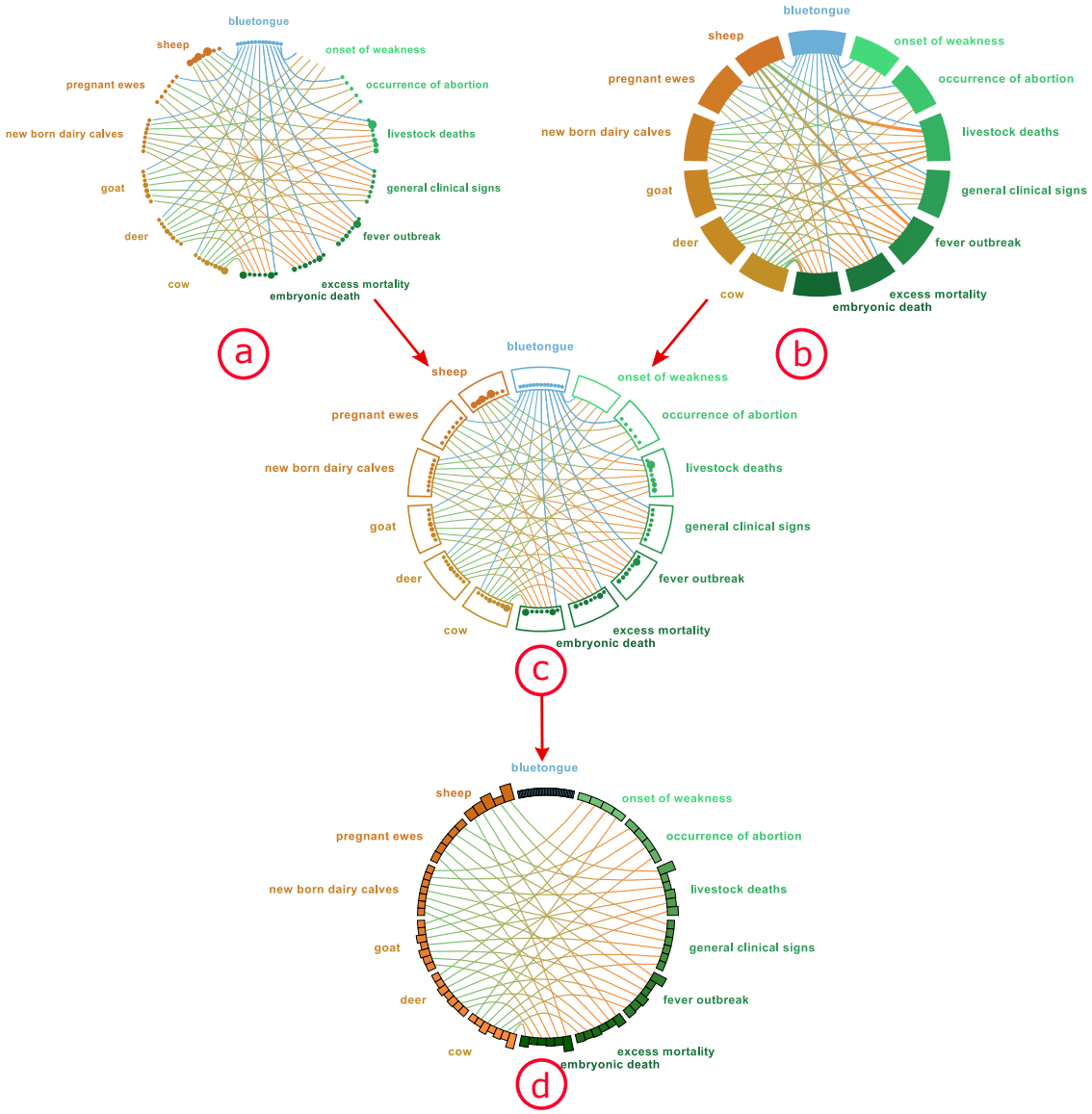


FIGURE 2.22 – Évolution de la visualisation de suggestions.

La Figure 2.23.a illustre un exemple dans lequel les étiquettes sont placées le long d'un arc formant un cercle de même centre que le cercle des objets mais avec un rayon plus important (e.g. [30,82]). Comme nous pouvons le constater, cette représentation est limitée à cause des problèmes de chevauchement qui peuvent apparaitre lorsque les étiquettes sont longues. Les experts ont aussi exprimé le fait qu'il était difficile de lire ces étiquettes à cause de la rotation.

La Figure 2.23.b illustre un exemple où les étiquettes sont arrangées de de façon radiale autour des objets (e.g. [18, 22,61,100]). Les experts ont également trouvé que cette représentation était difficile à appréhender dans la mesure où il était nécessaire de tourner la tête pour pouvoir les lire.

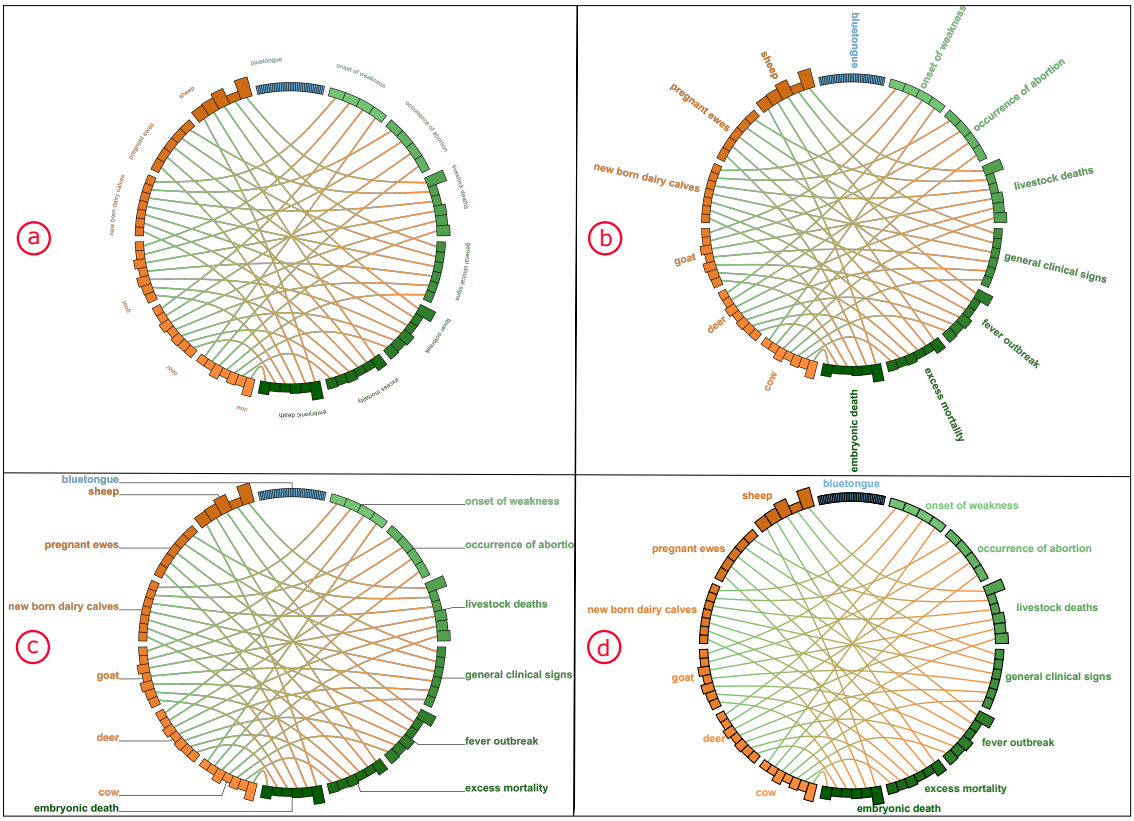


FIGURE 2.23 – Les différentes techniques d'étiquetage d'objets placés le long d'un cercle.

Une approche utilisant un affichage horizontal est proposée dans la Figure 2.23.c où des lignes permettent d'associer chaque étiquette avec son objet (e.g. [34]). Cette approche a l'avantage d'offrir plus de lisibilité des étiquettes mais surcharge la visualisation avec lignes. De plus, l'éloignement des objets et des étiquettes rend plus lent leur association.

Enfin la dernière approche positionne les étiquettes horizontalement au centre des objets (e.g. [10, 19, 87]) comme l'illustre la Figure 2.23.d. Cette approche a été retenue par les épidémiologistes qui ont particulièrement apprécié le fait que les textes soient lisibles et positionnés à côté des arcs.

2.5.3 Problème de chevauchements de nœuds

Lors de l'utilisation d'EPIDVIS, et plus particulièrement au moment de la fusion des nœuds dans le gestionnaire de mots-clés, il est apparu qu'un problème de chevauchement pouvait se présenter lorsqu'on repositionne un nœud fusionné au barycentre des nœuds qu'il contient. La Figure 2.24 illustre ce phénomène (voir les nœuds "*birds*", "*cow*" et "*ducks*"). Pour pallier ce problème, nous avons proposé

une nouvelle fonction de suppression de chevauchements de nœuds dans un dessin en une dimension. Cette contribution est détaillée dans le chapitre 3.

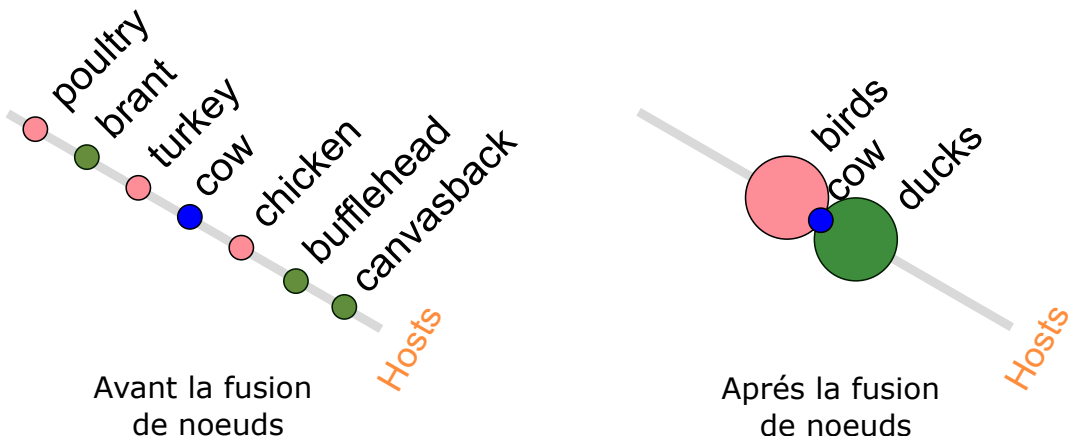


FIGURE 2.24 – Un exemple de chevauchement de nœuds sur un axe du gestionnaire de mots-clés. Les couleurs à gauche montrent les nœuds qui sont fusionnés dans l'axe de droite.

Chapitre

3

Suppression de chevauchement

Sommaire

3.1 Introduction

Dans le chapitre précédent, nous avons soulevé un problème technique provoqué par la fusion de termes dans le gestionnaire de mots-clés. Ce problème est le chevauchement des nœuds sur les axes. La Figure 3.1 montre un exemple du phénomène créé par plusieurs fusions successives de nœuds sur l'axe des espèces. Il peut bien entendu aussi se produire de la même façon sur les autres axes.

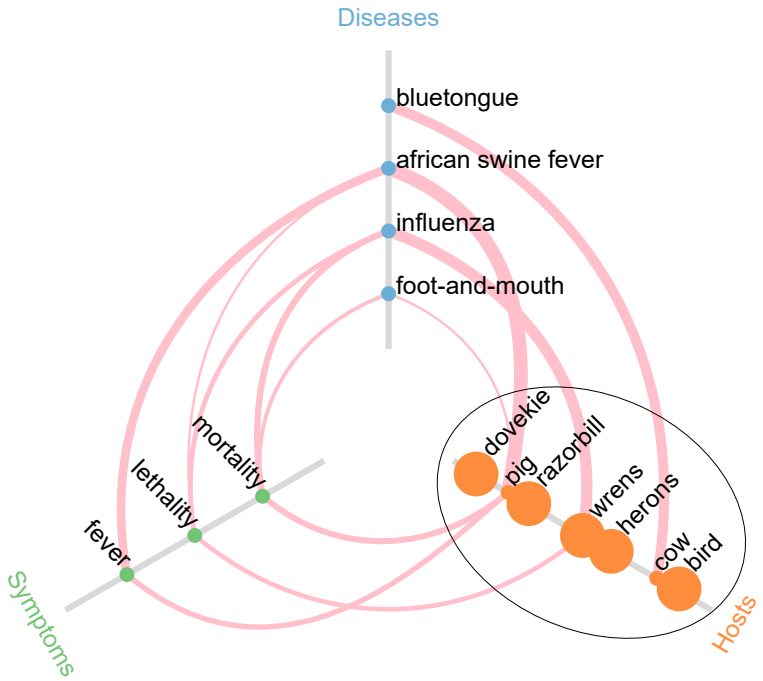


FIGURE 3.1 – Un exemple de problème de chevauchement dans l'axe des espèces.

En réalité, le problème des chevauchements des mots-clés le long des axes d'EPIDVIS peut être abordé de façon plus générale: c'est un problème de chevauchements de sommets positionnés sur une seule dimension. En particulier, il peut aussi concerner la visualisation de graphes produite à l'aide d'un algorithme de positionnement sur une dimension. Une représentation classique de graphe en 1D est connue sous le nom de diagrammes en arcs [97] : les nœuds sont positionnés uniformément le long d'un axe, et les arêtes sont dessinées comme des courbes au-dessus ou au-dessous de l'axe. La Figure 3.2 en montre un exemple représentant plusieurs suites de Farey [1]. Ici, les nœuds représentent des fractions, et les liens ayant la même couleur représentent une suite de Farey, i.e. une séquence particulière de fractions. Jusqu'à présent, la principale préoccupation des concepteurs de telles visualisations a été de trouver un ordre des nœuds qui minimise les longueurs des arêtes. Ce problème est appelé arrangement linéaire minimal [57, 74].

1. https://en.wikipedia.org/wiki/Farey_sequence

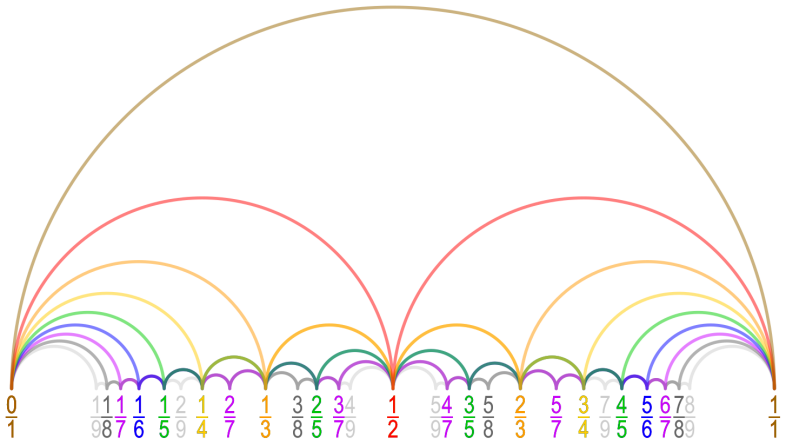


FIGURE 3.2 – Neuf suites de Farey représentées à l'aide d'un diagramme en arcs.

Une alternative au positionnement uniforme des nœuds le long d'un axe consiste à utiliser un algorithme de positionnement multidimentionnel, "Multidimensional Scaling" en anglais (MDS) [13]. Ce type d'algorithme part d'une matrice de dissimilarité entre des objets et tente de trouver un positionnement de ces objets minimisant la différence entre les valeurs de la matrice et les distances euclidiennes entre les positions. Dans le cadre du dessin de graphes, la matrice de dissimilarité des nœuds prise en entrée de l'algorithme correspond généralement aux longueurs des plus courts chemins entre les nœuds. L'algorithme tend ainsi à positionner les nœuds de telle sorte que les distances dans le dessin reflètent les longueurs des chemins dans le graphe. Le principal avantage de ce type de techniques est de regrouper les nœuds des parties denses du graphe, ce qui permet de mettre en évidence les longueurs des chemins et les structures communautaires, tout en montrant les symétries.

La bibliographie sur les techniques MDS permettant de dessiner des graphes inclut les dessins 2D [15, 43, 54, 76] et les dessins 1D [58, 59]. Lorsqu'il s'agit de graphes de grandes tailles, un inconvénient important de ces techniques est qu'elles produisent des dessins dans lesquels de nombreux nœuds se chevauchent, entraînant des problèmes d'encombrement visuel ("visual cluttering"). Plusieurs algorithmes ont été proposés pour traiter ce problème dans le cas des dessins de graphes en 2D [1, 32, 42, 47, 52, 68, 70]. Cependant, bien que la plupart d'entre eux puissent être adaptés au problème équivalent en 1D, leur temps d'exécution est relativement long (quadratique) et il est possible de trouver des alternatives plus efficaces, aussi bien en temps qu'en précision.

Le but de notre contribution [35] dans cette section est de proposer un algorithme permettant des repositionner des nœuds placés sur 1D afin de supprimer leurs chevauchements, tout en préservant autant que possible la configuration initiale du dessin. À notre connaissance, aucun algorithme n'a été proposé pour le dessin de graphe en 1D exclusivement. Notre contribution est double :

1. Nous proposons une définition formelle du problème de chevauchement des nœuds pour le dessin de graphes 1D. Il se compose de 4 critères ("requirements"). Bien que certains aient été abordés par des approches 2D, et pourraient être adaptés en 1D, le quatrième est nouveau et induit des propriétés intéressantes sur le dessin.
2. Nous proposons une méthode satisfaisant ces critères.

Nous illustrons aussi les résultats obtenus grâce à notre méthode avec deux études de cas et nous discutons au sujet des propriétés qu'elle induit.

Ce chapitre est organisé de la manière suivante. Nous définissons dans la section 3.2 les 4 critères à respecter pour la suppression des chevauchements de nœuds dans un graphe 1D. Ensuite, nous proposons un algorithme de complexité temporelle $O(|V|log(|V|)$ (où V est l'ensemble des nœuds) répondant à ces critères dans la section 3.3. Nous illustrons notre approche avec deux études de cas dans la section 3.4. Nous discutons quelques propriétés intéressantes induites par notre algorithme dans la section 3.5, et nous comparons notre technique avec des techniques alternatives dans la section 3.6. Comme dans les chapitres précédents, nous proposons une discussion pour conclure ce chapitre dans la section 3.7.

3.2 Définition du problème

Tout d'abord, nous partons d'un dessin initial dans lequel les nœuds sont positionnés le long d'un axe. Formellement, si V est l'ensemble de nœuds $\{v_1, v_2, \ldots, v_{|V|}\}$, le positionnement en 1D est donné par une fonction $p : V \to \mathbb{R}^+$.

Pour la suite, nous avons besoin d'une fonction représentant l'ordre des sommets dans le dessin initial. Nous définissons donc la fonction d'ordre $\sigma : V \to \{1, \ldots, |V|\}$ de la façon suivante : pour toute paire $(u, v) \in V^2$ avec $\sigma(u) + 1 = \sigma(v)$, $\nexists w \in V | p(u) < p(w) < p(v)$. S'il existe $(u, v) \in V^2$ tel que $p(u) = p(v)$, nous considérons un ordre total strict arbitraire sur les nœuds V pour satisfaire la contrainte suivante : $\forall (u, v) \in V^2, u \neq v \Rightarrow \sigma(u) \neq \sigma(v)$. σ^{-1} est la fonction inverse, $\sigma^{-1} : \{1, \ldots, |V|\} \to V$ tel que $\sigma^{-1}(i) = v \in V \Leftrightarrow \sigma(v) = i$.

Une fonction $s : V \to \mathbb{R}^+$ indique la taille des nœuds.

L'objectif ici est de trouver une fonction $f : V \to [0, l]$ qui supprime les chevauchements entre les nœuds, où l est la longueur du segment sur lequel nous voulons re-positionner les nœuds. Ce réarrangement doit permettre *(i)* d'améliorer la visualisation des éléments et *(ii)* préserver les principales caractéristiques du dessin initial (par exemple, l'ordre et la taille des nœuds). Ainsi, nous définissons un ensemble de 4 critères pour f afin de préserver la configuration globale du dessin initial :

Besoin 1. *Le dessin final doit utiliser de manière optimale la longueur du segment :*

— $f(\sigma^{-1}(1)) = 0 + s(\sigma^{-1}(1))/2$;
— $f(\sigma^{-1}(|V|)) = l - s(\sigma^{-1}(|V|))/2$.

Par exemple, si nous voulons utiliser toute la longueur de l'axe de la Figure 3.3.a représentant le positionnement initial, nous devons obtenir le positionnement de la Figure 3.3.b, dans lequel les nœuds "*dovekie*" et "*bird*" sont placés aux extrémités gauche et droite de l'axe.

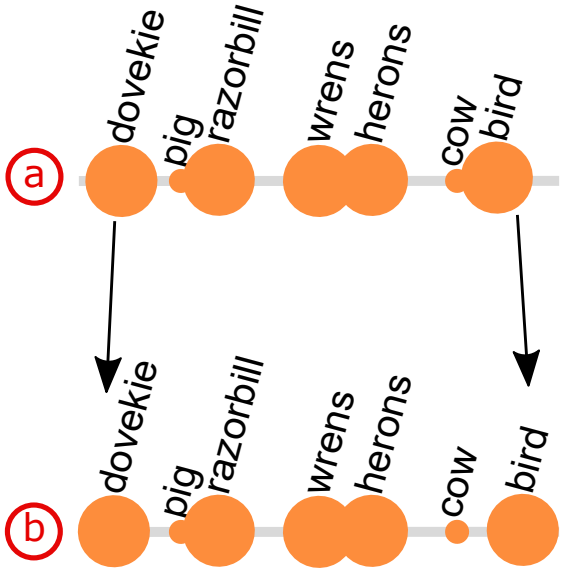


FIGURE 3.3 – Utilisation optimale de la longueur de l'axe de **(a)** dans **(b)**.

Besoin 2. *Le dessin final ne doit pas avoir de nœuds se chevauchant :*

$$\forall (u,v) \in V^2, |f(v) - f(u)| \geq s(u)/2 + s(v)/2$$

Par exemple, les nœuds qui se chevauchent sur la Figure 3.4.a doivent être séparés dans le positionnement final représenté par la Figure 3.4.b.

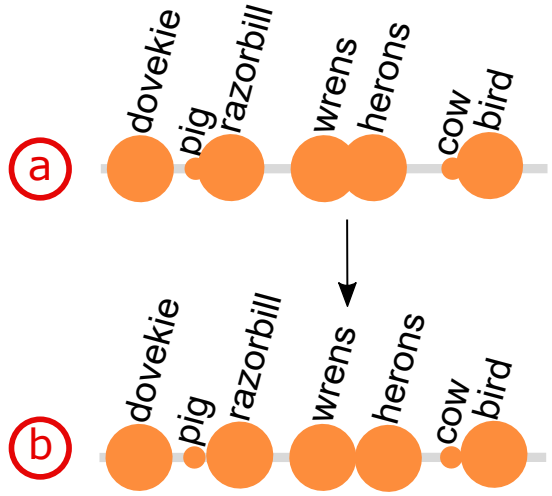


FIGURE 3.4 – Suppression des chevauchements des nœuds de **(a)** dans **(b)**

Besoin 3. *Le dessin final doit conserver l'ordre initial des nœuds. Comme les nœuds ne doivent pas se chevaucher, le critère peut être formalisé comme suit :*

$$\forall (u,v) \in V^2, \sigma(u) < \sigma(v) \Rightarrow f(u) < f(v)$$

Par exemple, l'ordre initial des nœuds "*wrens*" et "*herons*" dans la Figure 3.5.a ne doit pas être inversé dans le dessin final, comme c'est le cas dans la Figure 3.5.b.

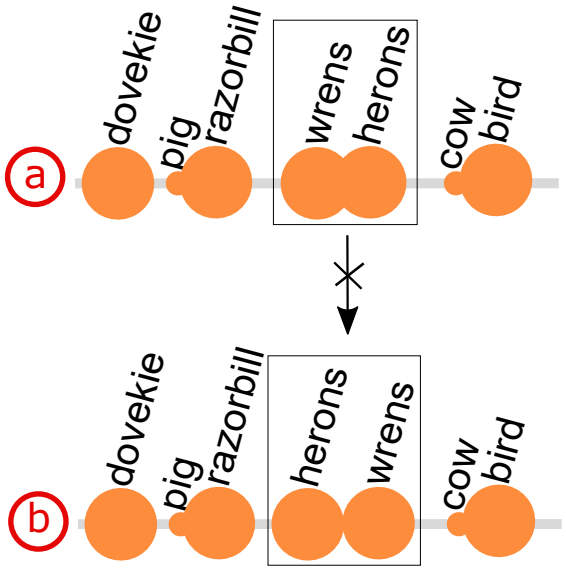


FIGURE 3.5 – Problème de changement de l'ordre des nœuds entre le positionnement initial **(a)** et le positionnement final **(b)**.

Besoin 4. *Le dessin final doit préserver les distances relatives entre les paires de nœuds consécutifs :* $\forall((u,v),(u',v')) \in V^2 \times V^2$ *tel que* $\sigma(u)+1 = \sigma(v)$ *et* $\sigma(u')+1 = \sigma(v')$,

$$p(v) - p(u) \geq p(v') - p(u') \Rightarrow$$
$$f(v) - \frac{s(v)}{2} - f(u) - \frac{s(u)}{2} \geq f(v') - \frac{s(v')}{2} - f(u') - \frac{s(u')}{2}$$

Par exemple, la distance entre les nœuds "herons" et "cow" dans la Figure 3.5.a est la plus grande des distances entre les paires de nœuds consécutifs. Elle doit donc rester la plus grande dans le dessin final de la Figure 3.5.b.

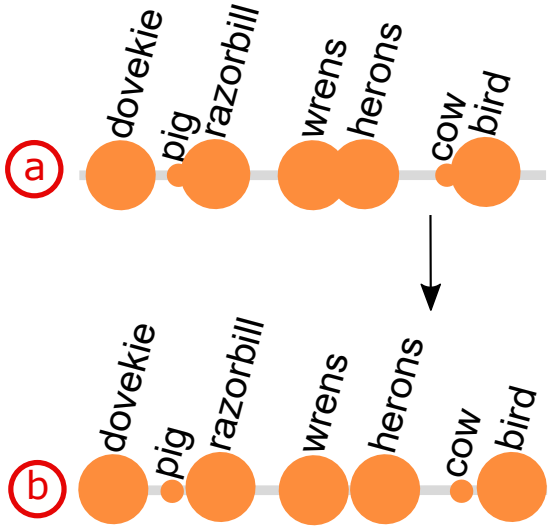


FIGURE 3.6 – Suppression des chevauchements de **(a)** avec préservation des distances relatives entre les nœuds dans **(b)**.

Remarque 1. *Supposons que* $\sum_{v \in V} s(v) \leq l$. *Dans ce cas, le problème n'a pas de solution, car la longueur du segment sur lequel les nœuds doivent être positionnés n'est pas assez grande pour contenir tous les nœuds sans chevauchement. Dans le cas où* $\sum_{v \in V} s(v) = l$, *les nœuds seront positionnés uniformément et bord à bord le long du segment en fonction de l'ordre* σ.

Remarque 2. *Il est essentiel de noter que le critère 4 définit la distance entre deux nœuds consécutifs* u *et* v *comme la distance entre leurs bordures respectives. Il pourrait aussi être défini comme la distance entre leurs centres. Les conséquences de ce choix sont discutées plus en détail dans la section 3.5.*

3.3 Notre proposition

Dans cette section, nous proposons une fonction qui satisfait les critères listés dans la section 3.2. Nous définissons d'abord p' comme une fonction de mise à l'échelle des valeurs de $p(v)$, pour $v \in V$, dans l'intervalle $[0, l - \sum_{x \in V} s(x)]$, avec $p_{min} = p(\sigma^{-1}(1))$ et $p_{max} = p(\sigma^{-1}(|V|))$:

$$p'(v) = \frac{p(v) - p_{min}}{p_{max} - p_{min}} \times (l - \sum_{x \in V} s(x))$$

D'un point de vue intuitif, cette fonction modifie les positions en ne considérant que la longueur du segment disponible une fois que tous les nœuds ont été placés sans se chevaucher, c-à-d, $l - \sum_{x \in V} s(x)$. La fonction p' assure une utilisation optimale de cette longueur disponible selon le critère 1.

Remarque 3. *Si $p(v) - p_{min}$ et $p_{max} - p_{min}$ sont positifs, $p'(v) \geq p'(u) \Leftrightarrow p(v) \geq p(u)$. Donc, p' conserve l'ordre fourni par p. De plus, $\forall (u, v) \in V^2, \sigma(u) < \sigma(v) \Rightarrow p'(u) \leq p'(v)$. Si $p(u) = p(v)$, alors $p'(u) = p'(v)$.*

La fonction de suppression des chevauchements des nœuds proposée, f, utilise la fonction p'. Elle considère la taille de chaque nœud de sorte que la position du centre d'un nœud v est donnée par $p'(v)$. Puis, elle ajoute le rayon de v et la taille de chaque nœud positionné avant v :

$$f(v) = p'(v) - \frac{s(v)}{2} + \sum_{i=1}^{\sigma(v)} s(\sigma^{-1}(i))$$

Remarque 4. *Nous avons déjà vu que $p(u) = p(v) \Rightarrow p'(u) = p'(v)$. Nous avons également mentionné que si $p(u) = p(v)$, nous considérons un ordre total strict arbitraire sur les nœuds de V pour définir σ. Par exemple, considérons u et v tel que $p(u) = p(v)$ et $\sigma(v) = \sigma(u) + 1$. Dans ce cas, $f(v) = f(u) + s(u)/2 + s(v)/2$, c-à-d, les nœuds sont positionnés côte à côte sans chevauchement.*

Théoréme 1. *La fonction f satisfait les 4 critères définis ci-dessus.*

Démonstration pour le critère 1 (Utilisation optimale de la longueur du segment). On définit les nœuds v_1 et $v_{|V|}$ tel que $v_1 = \sigma^{-1}(1)$ et $v_{|V|} = \sigma^{-1}(|V|)$.

$$f(v_1) = p'(v_1) - \frac{s(v_1)}{2} + \sum_{i=1}^{\sigma(v_1)} s(\sigma^{-1}(i)) \tag{3.1}$$

$$= 0 - \frac{s(v_1)}{2} + s(v_1) \tag{3.2}$$

$$= \frac{s(v_1)}{2} \tag{3.3}$$

$$f(v_{|V|}) = p'(v_{|V|}) - \frac{s(v_{|V|})}{2} + \sum_{i=1}^{\sigma(v_{|V|})} s(\sigma^{-1}(i)) \tag{3.4}$$

$$= l - \sum_{v \in V} s(v) - \frac{s(v_{|V|})}{2} + \sum_{v \in V} s(v) \tag{3.5}$$

$$= l - \frac{s(v_{|V|})}{2} \tag{3.6}$$

Nous obtenons donc les valeurs requises pour les positions du premier et du dernier nœud.

□

Démonstration pour le critère 2 (Pas de chevauchement). Nous développons d'abord la définition de $f(v)$ comme suit, étant donné que $(u, v) \in V^2$ et $\sigma(u) < \sigma(v)$:

$$f(v) = p'(v) - \frac{s(v)}{2} + \sum_{i=1}^{\sigma(v)} s(\sigma^{-1}(i)) \tag{3.7}$$

$$= p'(u) + p'(v) - p'(u) - \frac{s(v)}{2} + \sum_{i=1}^{\sigma(v)} s(\sigma^{-1}(i)) \tag{3.8}$$

Nous isolons de $\sum_{i=1}^{\sigma(v)} s(\sigma^{-1}(i))$ les nœuds ordonnés jusqu'à u à travers la fonction d'ordre σ :

$$\begin{aligned} f(v) = {} & p'(u) + p'(v) - p'(u) - \frac{s(v)}{2} \\ & + \sum_{i=1}^{\sigma(u)} s(\sigma^{-1}(i)) + \sum_{i=\sigma(u)+1}^{\sigma(v)} s(\sigma^{-1}(i)) \end{aligned} \tag{3.9}$$

En ajoutant $\frac{s(u)}{2} - \frac{s(u)}{2}$, nous identifions la définition de $f(u)$, et nous réécrivons $f(v)$ en fonction de $f(u)$:

$$\begin{aligned} f(v) = {} & \left(p'(u) - \frac{s(u)}{2} + \sum_{i=1}^{\sigma(u)} s(\sigma^{-1}(i)) \right) \\ & + p'(v) - p'(u) - \frac{s(v)}{2} + \sum_{i=\sigma(u)+1}^{\sigma(v)} s(\sigma^{-1}(i)) + \frac{s(u)}{2} \end{aligned} \tag{3.10}$$

$$= f(u) + p'(v) - p'(u) - \frac{s(v)}{2} + \sum_{i=\sigma(u)+1}^{\sigma(v)} s(\sigma^{-1}(i)) + \frac{s(u)}{2} \tag{3.11}$$

Nous extrayons de $\sum_{i=\sigma(u)+1}^{\sigma(v)} s(\sigma^{-1}(i))$ la taille de v. Nous obtenons ainsi une expression de $f(v)$ comme somme de termes positifs, à condition que $p'(v) - p'(u)$ soit positif (Cf. Remarque 3).

$$f(v) = f(u) + p'(v) - p'(u) + \frac{s(v)}{2} + \sum_{i=\sigma(u)+1}^{\sigma(v)-1} s(\sigma^{-1}(i)) + \frac{s(u)}{2} \tag{3.12}$$

ce qui équivaut à :

$$f(v) - f(u) = p'(v) - p'(u) + \frac{s(v)}{2} + \sum_{i=\sigma(u)+1}^{\sigma(v)-1} s(\sigma^{-1}(i)) + \frac{s(u)}{2} \tag{3.13}$$

Pour deux nœuds $(u, v) \in V^2$ tel que $\sigma(u) < \sigma(v)$, $p'(v) - p'(u) \geq 0$ (Cf. Remarque 3). De plus, la somme des tailles des nœuds $\sum_{i=1}^{\sigma(v)} s(\sigma^{-1}(i))$ est positive. Par conséquent :

$$f(v) - f(u) \geq \frac{s(u)}{2} + \frac{s(v)}{2} \tag{3.14}$$

et par extension :

$$\forall (u, v) \in V^2, |f(v) - f(u)| \geq s(u)/2 + s(v)/2. \tag{3.15}$$

□

Démonstration pour le critère 3 (Préserver l'ordre initial). La preuve du critère 2 nous montre que pour deux nœuds $(u, v) \in V^2$ tel que $\sigma(u) < \sigma(v)$, on a $f(v) - f(u) \geq s(u)/2 + s(v)/2$. Puisque $s(u)/2 + s(v)/2$ est strictement positif, $f(v) - f(u)$ l'est aussi. Par conséquent, $\forall (u, v) \in V^2, \sigma(u) < \sigma(v) \Rightarrow f(u) < f(v)$.

□

Pour prouver que la fonction f satisfait le critère 4, nous avons besoin du lemme suivant.

Lemme 1. *La distance entre deux nœuds consécutifs u et v dans le dessin final est égal à la différence $p'(v) - p'(u)$. Plus formellement, soient deux nœuds $(u, v) \in V^2$:*

$$\sigma(u) + 1 = \sigma(v) \Rightarrow f(v) - \frac{s(v)}{2} - (f(u) + \frac{s(u)}{2}) = p'(v) - p'(u)$$

Démonstration. Notons D comme étant la somme $f(v) - \frac{s(v)}{2} - (f(u) + \frac{s(u)}{2})$. Nous développons D en appliquant la définition de $f(v)$:

$$D = p'(v) - \frac{s(v)}{2} + \sum_{i=1}^{\sigma(v)} s(\sigma^{-1}(i)) - \frac{s(v)}{2} - (f(u) + \frac{s(u)}{2}) \tag{3.16}$$

$$= p'(v) - s(v) + \sum_{i=1}^{\sigma(v)} s(\sigma^{-1}(i)) - (f(u) + \frac{s(u)}{2}) \tag{3.17}$$

Ensuite, en appliquant la définition de $f(u) = p'(u) - \frac{s(u)}{2} + \sum_{i=1}^{\sigma(u)} s(\sigma^{-1}(i))$:

$$D = p'(v) - s(v) + \sum_{i=1}^{\sigma(v)} s(\sigma^{-1}(i)) - p'(u) - \sum_{i=1}^{\sigma(u)} s(\sigma^{-1}(i)) \tag{3.18}$$

Comme $\sigma(v) = \sigma(u) + 1$, $\sum_{i=1}^{\sigma(v)} s(\sigma^{-1}(i)) - \sum_{i=1}^{\sigma(u)} s(\sigma^{-1}(i)) = s(v)$, l'équation précédente peut être simplifiée pour prouver le lemme 1 :

$$f(v) - \frac{s(v)}{2} - (f(u) + \frac{s(u)}{2}) = p'(v) - s(v) - p'(u) + s(v) \tag{3.19}$$

$$= p'(v) - p'(u) \tag{3.20}$$

□

Démonstration pour le critère 4 (Préserver les distances relatives). En appliquant le lemme 1, nous devons prouver que :

$$p(v) - p(u) \geq p(v') - p(u') \Rightarrow p'(v) - p'(u) \geq p'(v') - p'(u')$$

Nous développons d'abord $p'(v) - p'(u)$ en appliquant la définition de p :

$$p'(v) - p'(u) = \frac{p(v) - p_{min}}{p_{max} - p_{min}} \times C - \frac{p(u) - p_{min}}{p_{max} - p_{min}} \times C \tag{3.21}$$

où $C = (l - \sum_{x \in V} s(x))$, et $C \geq 0$ (Cf. Remarque 1). Ainsi :

$$p'(v) - p'(u) = C \left(\frac{p(v) - p(u)}{p_{max} - p_{min}} \right) \tag{3.22}$$

$$= \frac{C}{p_{max} - p_{min}} \times (p(v) - p(u)) \tag{3.23}$$

Puisque $C' = \frac{C}{p_{max} - p_{min}}$ est positif, nous obtenons :

$$p(v) - p(u) \geq p(v') - p(u') \Rightarrow C' \left(p(v) - p(u)\right) \geq C' \left(p(v') - p(u')\right)$$
$$p(v) - p(u) \geq p(v') - p(u') \Rightarrow p'(v) - p'(u) \geq p'(v') - p'(u')$$

□

Théoréme 2. *Étant donné un graphe positionné sur 1D, un dessin sans chevauchement peut être calculé en temps O(|V|log(|V|)).*

Démonstration. Soit p une fonction donnée contenant les positions initiales des nœuds, l'ordre σ se calcule en temps $O(|V|log(|V|))$. La valeur de f pour chaque nœud est calculée en $O(|V|)$ en procédant selon l'ordre σ, et en stockage les tailles cumulées des nœuds précédemment traités dans une variable. □

3.4 Études de cas

3.4.1 *Les Misérables*

Cette première étude de cas est effectuée sur le jeu de données *Les Misérables* [56]. Chaque nœud représente un personnage du roman de Victor Hugo. Une arête reliant une paire de nœuds signifie que les deux personnages sont cités ensemble dans au moins un chapitre du livre. La taille d'un nœud correspond au nombre de co-occurrences du personnage dans les différents chapitres. Elle correspond donc au degré du nœud dans le graphe.

La Figure 3.7 montre deux représentations 1D distinctes de ce jeu de données. La Figure 3.7 (a) représente le dessin initial obtenu en appliquant l'approche MDS [28]. Ce dessin contient des chevauchements. La Figure 3.7 (b) montre les mêmes données sans chevauchement, obtenu en appliquant notre méthode sur un segment de même longueur que le segment initial. À partir de cette figure, nous pouvons voir que des informations ont été mises en évidence grâce à notre méthode. Par exemple, la partie la plus à gauche du dessin initial semble être composée d'un seul nœud à partir duquel une arête sort. Ensuite, cette arête est divisée en 4 arêtes distinctes reliant cet élément aux 4 nœuds les plus proches. Le deuxième dessin obtenu en supprimant les chevauchements montre que le cercle bleu le plus à gauche du dessin initial correspond en réalité à plusieurs nœuds. Il permet à l'utilisateur d'identifier comment chaque nœud (c-à-d personnage) est connecté aux autres éléments.

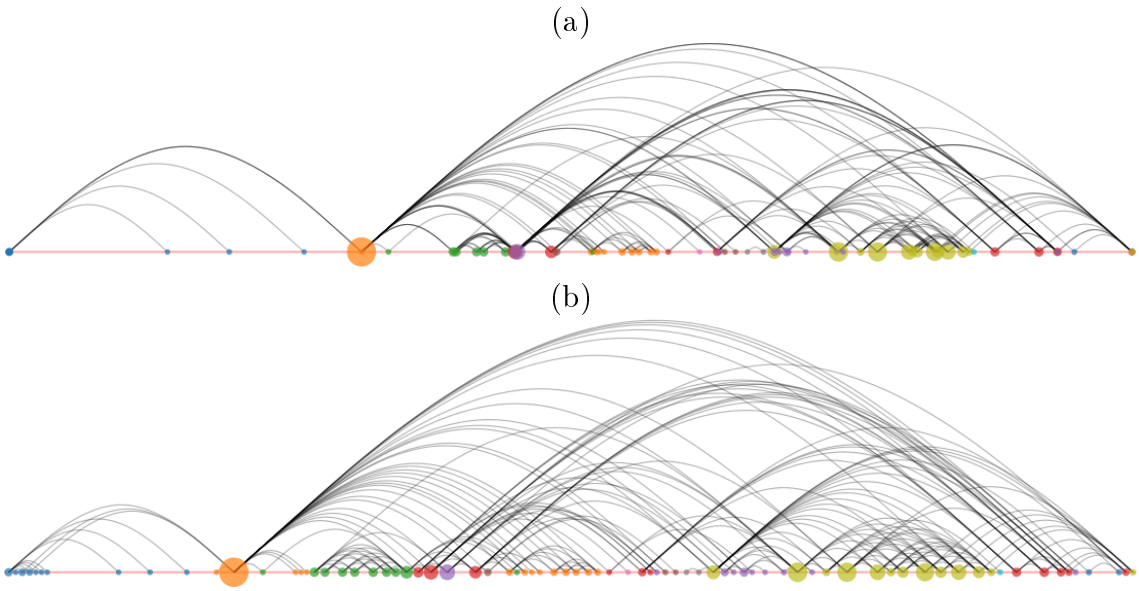


FIGURE 3.7 – Deux dessins 1D représentant le jeu de données *Les Misérables* : **(a)** dessin initial contenant des chevauchements, **(b)** dessin final après suppression des chevauchements.

La Figure 3.8 montre une sous-partie des visualisations précédentes. La Figure 3.8 (a) et la Figure 3.8 (b) correspondent respectivement au même sous-ensemble de nœuds avec le dessin initial et le dessin après suppression des chevauchements (les arêtes ont été supprimées pour mieux se concentrer sur le placement de nœuds).

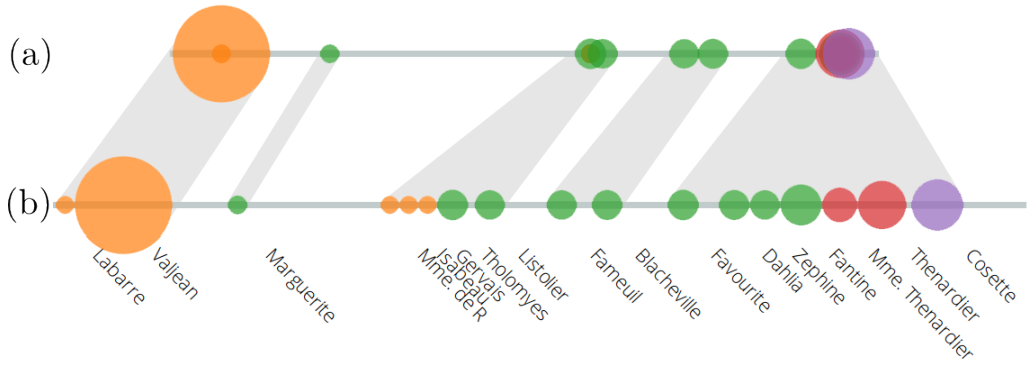


FIGURE 3.8 – Sous-partie du jeu de données *Les Misérables* : **(a)** dessin initial, **(b)** dessin après suppression des chevauchements. Les correspondances entre les clusters de nœuds dans les deux dessins sont surlignées en gris clair.

Dans la Figure 3.8 (a), beaucoup de nœuds se chevauchent. On peut observer la présence de différents clusters mais les chevauchements empêchent toute analyse au niveau des nœuds. Comme le montre la Figure 3.8 (b), notre méthode joue son rôle comme prévu. Elle permet de voir tous les nœuds, même ceux qui faisaient initialement partie d'un ensemble se chevauchant. Elle offre aussi une meilleure lisibilité des données à la plus fine granularité. Par exemple, nous pouvons considérer le cluster le plus à droite de la partie supérieure de la vue. Il est composé de plusieurs nœuds mais on ne peut pas savoir leur nombre sur la première figure. Dans la deuxième figure, chaque nœud est distinguable et on peut voir quels sont les personnages du cluster. En outre, la figure montre également que certains nœuds tels que *Favorite* qui apparaissent initialement très proches d'autres nœuds sont en réalités légèrement éloignés.

Nous sommes convaincus que le dessin proposé permet d'analyser plus facilement les distances entre paires de nœuds consécutifs. Par exemple, on peut facilement observer dans la seconde figure que la distance entre les personnages *Valjean* et *Marguerite* (représentée comme l'espace vide entre les deux nœuds) est environ 1/3 de la distance entre *Marguerite* et *Madame de R.* Cette observation est plus difficile à obtenir à partir du dessin initial, en particulier lorsque les nœuds ont des tailles différentes, car elle nécessite de considérer la distance entre les centres des nœuds (voir la section 3.5 pour plus de détails au sujet du critère 4).

Nous remarquons également que les nœuds apparaissant initialement seuls, tels que *Marguerite*, apparaissent également seuls dans le dessin final.

3.4.2 Réseau de co-auteurs de PacificVis

La deuxième étude de cas est réalisée sur un réseau de co-auteurs de *PacificVis* extrait de DBLP. Chaque nœud du graphe représente un auteur ayant publié au moins 3 articles dans le symposium entre 2008 et 2016. Il y a une arête entre deux nœuds si les auteurs correspondants ont publié au moins un article ensemble. Nous avons également extrait pour chaque auteur un vecteur contenant le nombre d'articles publiés par année. La taille de chaque nœud correspond au nombre total d'articles publiés par l'auteur durant la période.

Rufiange et McGuffin [78] proposent 5 stratégies pour visualiser des graphes dynamiques (Cf. Figure 2 de leur article). Il est important de noter que le concept de graphe dynamique développé dans l'article couvre à la fois l'évolution topologique des graphes (ajout/suppression de nœuds/liens au cours du temps), et l'évolution d'attributs numériques associés aux nœuds/liens. La 4ème stratégie proposée par Rufiange et McGuffin consiste à visualiser l'axe du temps perpendiculairement à un diagramme en arcs. Dans leur exemple, l'axe vertical permet de représenter le diagramme en arcs et l'axe horizontal permet de représenter l'évolution d'un attribut numérique sur les nœuds. Dans cette section, nous nous concentrons sur cette stratégie en traitant l'évolution d'un attribut numérique comme dans l'exemple de l'article.

Le réseau de co-auteurs est représenté sous la forme d'un diagramme en arcs vertical (Cf. Figure 3.9). Les nœuds sont positionnés en appliquant l'approche MDS [28] suivie de notre méthode pour supprimer les chevauchements des nœuds. Le vecteur contenant le nombre d'articles publiés chaque année par un auteur est représenté par un graphique silhouette horizontal ("silhouette graph") [46] à côté de chaque nœud. L'échelle verticale uniforme des graphiques silhouette et l'alignement vertical des années assurent que les nombres de publications sont faciles à comparer d'un auteur à un autre. Les lignes grises verticales représentent les années. Chaque composante connexe du réseau a été traitée séparément des autres. Les résultats correspondants ont été placés verticalement sur le même dessin dans un ordre arbitraire. Les couleurs sont fournies par la configuration par défaut de la bibliothèque D3 [14].

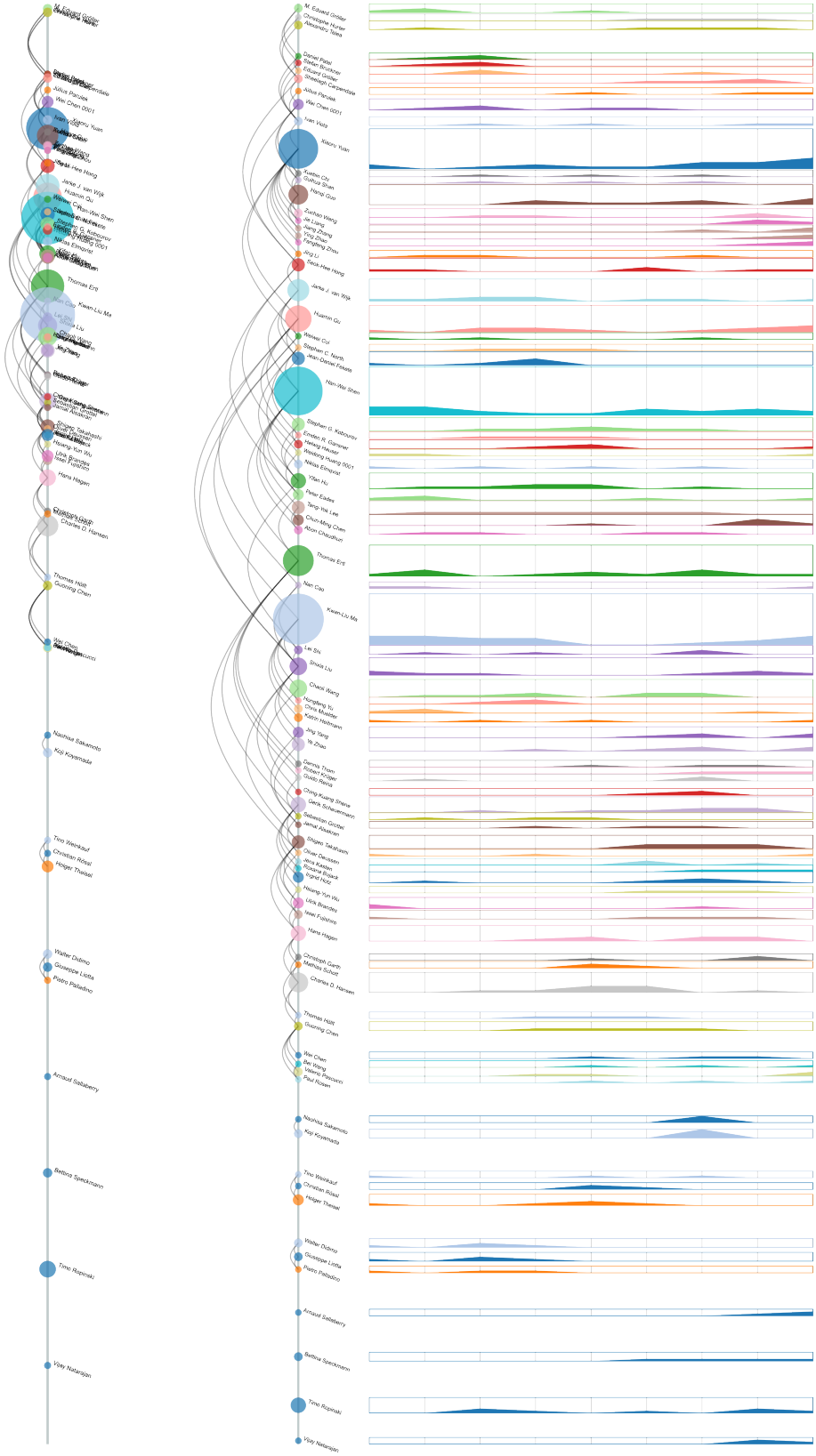


FIGURE 3.9 – Réseau de co-auteurs de *PacificVis* extrait de DBLP, avec chevauchements (à gauche), sans chevauchement (au centre), et graphique silhouette (à droite).

La Figure 3.9 montre que le réseau est constitué de 8 composantes connexes. La première (à partir du haut) contient une très grande partie de l'ensemble de nœuds. Il est suivi d'une composante de deux éléments, deux composantes de trois éléments, et quatre singletons.

L'algorithme de dessin de graphe utilisé (MDS + suppression des chevauchements) permet de mettre en évidence l'importance des relations entre les nœuds (ici, elles sont liées au nombre de publications pendant la période 2008-2016), ainsi que certaines communautés au sein des nœuds. La partie droite de la visualisation permet d'obtenir facilement des informations sur chaque nœud en fonction de la dimension temporelle. Dans ce qui suit, nous discutons des avantages du dessin de graphe 1D proposé dans un tel contexte.

La Figure 3.10 montre une sous-partie du réseau de co-auteurs dans laquelle trois sous-ensembles d'auteurs (notés $S1$, $S2$ et $S3$) interagissent de manière intéressante. La distance entre les éléments de chaque sous-ensemble montre qu'ils font partie de petites communautés distinctes dans le réseau. Il est intéressant de noter que $S2$ joue un rôle de pivot en entre les communautés $S1$ et $S3$. En effet, les auteurs de $S1$ et $S3$ n'ont pas de liens directs (c-à-d de publications ensemble) mais ils sont connectés à un ou deux auteurs de $S2$. Il est également intéressant de noter que les auteurs d'une même communauté ont un graphique silhouette, donc une évolution de leur nombre de publications, assez similaire.

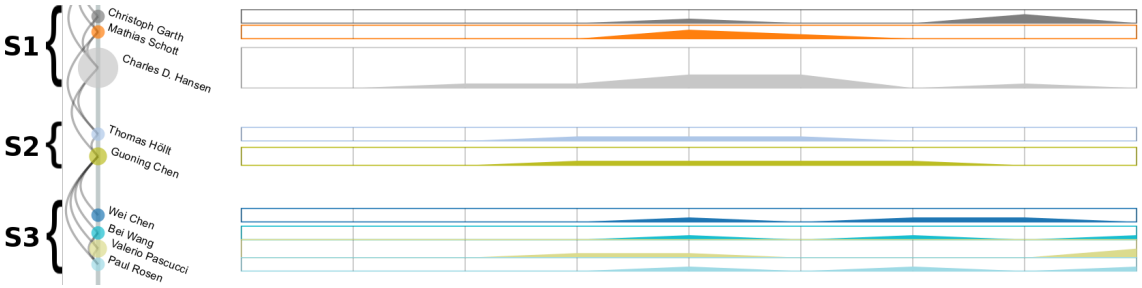


FIGURE 3.10 – Sous-partie du réseau de co-auteurs de *PacificVis*.

Un aspect important de la stratégie visuelle employée (1D pour le graphe, 1D pour l'attribut numérique) est qu'elle permet une bonne lisibilité des distributions des nombres de publications par auteur ainsi qu'un moyen facile de les comparer. La Figure 3.11 fournit une sélection d'auteurs avec des nombres totaux de publications assez similaires (tailles de nœuds assez proches). On peut observer différents profils dans l'évolution des nombres de publications par année. Prenons l'exemple de la Figure 3.11 (a), il correspond à un auteur dont le nombre de publications à *PacificVis* présente une régularité pendant la majeure partie de la période considérée. Ce profil de distribution de publications contraste avec le profil plus sporadique fourni par la Figure 3.11 (b) où le nombre total de publications est partagé entre deux éditions de la conférence. Le troisième auteur, Figure 3.11 (c), montre un profil de progression où le nombre de publications a débuté et augmenté au cours des deux dernières années.

La Figure 3.11 (d) montre un cas particulier où toutes les publications ont eu lieu la même année. Ce cas est facilement remarquable car une des valeurs représentées dans le diagramme silhouette atteint la bordure supérieure du cadre.

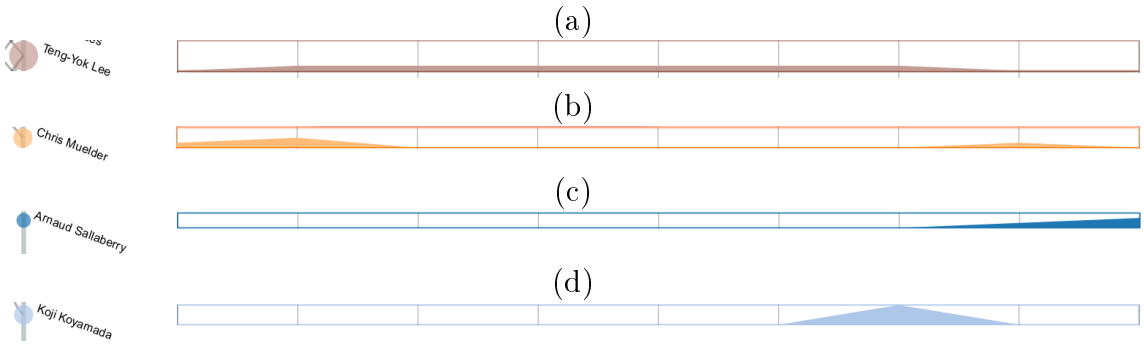


FIGURE 3.11 – Divers auteurs du réseau de co-auteurs de *PacificVis* présentant différents profils de publication au cours des ans.

D'un point de vue plus général, nous pouvons remarquer que la visualisation obtenue en alignant verticalement les graphiques silhouette facilite considérablement la comparaison des profils de publication dans le temps, et la comparaison des nombres de publications pour une année donnée. Cette aide à la comparaison des éléments du réseau serait beaucoup plus difficile à obtenir avec les 4 autres stratégies de représentation visuelle proposées par Rufiange et McGuffin [78] qui sont toutes basées sur des dessins en 2D. Il est intéressant de noter qu'aucun d'entre eux n'offre une telle facilité d'analyse lorsque l'utilisateur doit comparer visuellement le nombre de publications au cours des années. Cette étude de cas montre comment un dessin de graphe en 1D peut être une alternatife pertinente au dessin de graphe en 2D en permettant l'accomplissement plus efficace de certaines tâches, par exemple la comparaison de nœuds associés à des données temporelles [80].

Cet avantage est principalement dû au fait que l'utilisation d'un dessin du graphe en 1D pour positionner les nœuds permet à la représentation visuelle d'exploiter pleinement une dimension supplémentaire pour représenter l'attribut dynamique du graphe (ici la distribution des publications dans le temps). Habituellement, l'utilisation d'un dessin de graphe en 1D plutôt qu'en 2D implique un coût important au niveau de la visualisation de la structure des données dû au : **(1) manque de lisibilité**, l'utilisation d'une approche MDS pour placer les nœuds produit des nœuds qui se chevauchent; **(2) manque d'informations**, c'est le cas lorsque l'on utilise une approche d'arrangement linéaire minimal dans laquelle les nœuds sont placés uniformément sur la longueur du segment, ce qui ne permet pas de mettre en évidence les communautés. Cette étude de cas montre que le coût d'utilisation d'un dessin de graphe en 1D plutôt qu'en 2D est réduite en exploitant notre approche car elle supprime les chevauchements des nœuds et met en relief des propriétés structurelles des données initiales.

3.5 Discussion

3.5.1 Proximité des nœuds

Les techniques existantes pour supprimer les chevauchements de nœuds dans les graphes en 2D (e.g. [32,42]) déplacent les nœuds du dessin initial jusqu'à ce qu'il n'y ait plus de chevauchement dans le dessin (voir la section 3.6 pour plus de détails). Par conséquent, les distances entre les paires de nœuds dans le dessin final dépendent du chevauchement des nœuds dans le dessin initial, et le critère 4 n'est pas toujours satisfait. C'est différent avec notre méthode. La Figure 3.12 montre un exemple de graphe composé de trois nœuds u, v et w avec deux dessins obtenus en diminuant la longueur du segment l à l'. La distance entre les nœuds consécutifs (c-à-d la longueur entre leurs bordures les plus proches) diminue linéairement lorsque la longueur du segment diminue. En d'autres termes, les rapports des distances restent identiques : $d_1/d_2 = d'_1/d'_2$. Les approches en 2D ne permettent pas à l'utilisateur d'évaluer précisément la distance réelle entre deux nœuds. Au contraire, la méthode que nous proposons conserve toujours les distances relatives exactes en fonction de dessin initial.

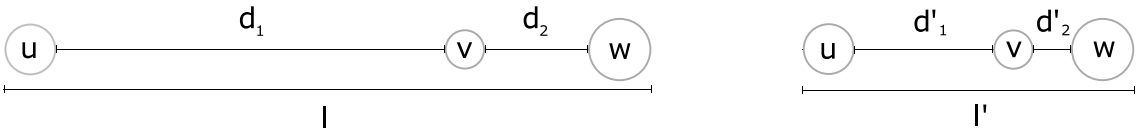


FIGURE 3.12 – Deux dessins obtenus en diminuant la longueur du segment de l à l'.

3.5.2 Perception de distances

En plus de contenir des nœuds qui se chevauchent, le dessin initial donné par la fonction p dans la section 3.2 peut également avoir l'inconvénient de conduire les utilisateurs à interpréter à tort les distances entre les nœuds. Considérons l'exemple fourni dans la Figure 3.13, trois nœuds u, v et w sont représentés dans deux dessins en 1D distincts : le dessin initial fourni par la fonction p, et la fonction f proposée dans cet article.

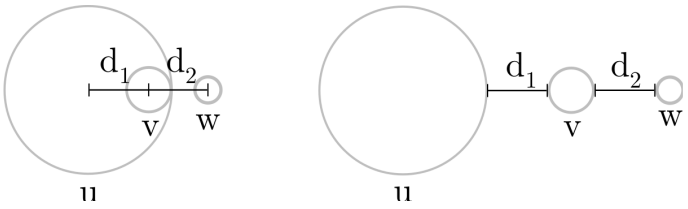


FIGURE 3.13 – Dessin initial à gauche et notre dessin à droite.

Dans le dessin initial, on peut remarquer que le nœud v est perçu comme inclut dans le nœud u, ce qui peut provoquer des erreurs d'interprétation au sujet de la

distance réelle qui les sépare : u et v semblent plus proches que v et w. Toutefois, si l'on regarde plus attentivement les distances entre les centres des nœuds sans tenir compte de leur taille, on peut facilement observer qu'il s'agit d'une erreur de perception : v et w sont actuellement à la même distance que u et v. Ce biais de perception est directement lié à la loi de la continuité issue de la théorie de la psychologie de la forme ("Gestalt psychology").

Un des objectifs de notre approche est de fournir une solution simple et efficace à ce problème en produisant un dessin dans lequel la distance entre deux nœuds correspond à la distance entre leurs bordures les plus proches. La figure de droite illustre ce phénomène : les nœuds v et w semblent être aussi proches que u et v, ce qui reflète correctement les données. Le critère 4 permet donc une compréhension plus rapide et plus pertinente des données.

3.6 Techniques alternatives

De nombreuses approches ont été proposées pour supprimer les chevauchements de nœuds dans les dessins de graphes en 2D. La plupart d'entre elles peuvent être appliquées aux dessins en 1D [1,32,47,52,64,68,70]. Elles appartiennent toutes à deux catégories principales : *Force Scan Algorithms* (FSA) et *Constraint Optimization based Approches* (COA). Alors que leurs versions en 1D répondent au 2ème critère (pas de chevauchement), il n'est pas évident qu'elles satisfassent les autres. Dans cette section, nous donnons une brève description de ces algorithmes ainsi que de leur aptitude à satisfaire les critères et leur complexité en temps de calcul.

L'algorithme FSA, introduit par Misue *et al.* dans [70], est une approche majeure de la première catégorie. L'idée est d'appliquer une fonction de force f aux nœuds qui se chevauchent : si un nœud u se chevauche avec un autre nœud v, alors $f(u, v)$ repousse v de u. Les auteurs proposent un modèle de force par répulsion et un algorithme pour appliquer ce modèle afin de supprimer les chevauchements de nœuds (*Push Force Scan Algorithm*). Les forces pour tous les nœuds sont d'abord appliquées horizontalement, puis verticalement, de sorte que l'algorithme peut être utilisé sur un dessin en 1D. Les auteurs proposent également un second modèle de force permettant à la fois d'attirer et de repousser les nœuds : (*Push-Pull Force Scan Algorithm*). Le dessin est dans ce cas plus compact. Dans [47], Hayashi *et al.* prouvent que le problème de trouver un dessin d'aire minimale tout en supprimant les chevauchements et en préservant l'ordre orthogonal des nœuds est NP-complet. Ensuite, ils proposent un algorithme heuristique basé sur la même idée que FSA qui positionne les nœuds dans une zone plus petite. Un autre travail de Li *et al.* [64] propose aussi deux autres variantes de FSA qui donnent une plus petite surface. Huang *et al.* introduisent l'algorithme *Force-Transfer Algorithm* (FTA) dans [52]. L'algorithme commence avec un nœud de départ et trouve le groupe de nœuds chevauchants contenant le nœud de départ sur une dimension (horizontale ou verticale). Puis, un modèle de force est appliqué pour supprimer les chevauchements dans ce

cluster. Certains nœuds peuvent alors se chevaucher avec des nœuds en dehors du cluster de sorte qu'un nouveau cluster est formé dynamiquement. Le processus se termine lorsque le cluster de chaque nœud est vide. Le même processus est ensuite appliqué à l'autre dimension. Abe *et al.* affinent cet algorithme dans [1]. Le dessin issu de FTA est plus compact que celui issu de FSA. Les algorithmes FSA et FTA préservent l'ordre orthogonal, sauf celui présenté dans [64]. Ainsi, la version en 1D répond au 3ème critère. L'aire du dessin finale n'est pas délimité de sorte que l'algorithme ne répond pas au premier critère. Comme les tailles des nœuds ne sont pas considérées, le 4ème critère n'est pas satisfait. Ils ont une complexité de $O(|V|^2)$ pour les scans horizontaux et verticaux, donc s'ils sont appliqués sur une seule dimension, ils restent en $O(|V|^2)$.

Les approches de type COA se focalisent sur la minimisation d'une fonction objective tout en préservant un ensemble de contraintes. La fonction objective représente le mouvement de nœuds. Les contraintes garantissent que les nœuds ne se chevauchent pas. La suppression des chevauchements des nœuds est ainsi traitée comme un problème d'optimisation par contraintes. Dans [32], Dwyer *et al.* proposent un algorithme pour trouver les contraintes de séparation des nœuds, et un autre pour résoudre le problème d'optimisation correspondant. Cet algorithme peut être appliqué au dessin de graphe en 1D. Il ne répond pas aux critères 1 et 4. Même si la fonction objective assure de petites modifications des positions des nœuds, elle ne garantit pas que l'ordre orthogonal initial soit préservé (critère 3). Un autre travail proposé par Marriott *et al.* [68] décrit quatre approches de type COA. Aucune d'entre elles ne répond aux critères 1, 3 et 4. Les approches COA nécessitent de trouver d'abord les contraintes, puis de supprimer les chevauchements. Dans [32], la complexité pour trouver les contraintes de séparation est $O(|V|log|V|)$. L'algorithme de satisfaction de contraintes s'exécute en $O((|V| \cdot |C|)log|C|)$ où C est l'ensemble des contraintes (le nombre de ces contraintes est $O(|V|)$) [33]. Dans [68], la complexité pour trouver une solution sans chevauchement est $O(|V|^2 \cdot log(|V|))$ pour la première approche basée sur une mise à l'échelle uniforme. D'autres approches décrites dans le document prennent plus de temps.

La table 3.1 résume les propriétés des différentes alternatives, et leurs complexités en temps. On peut observer que notre algorithme est le plus rapide et le seul à satisfaire les 4 critères. Ceci est dû au fait qu'il est le seul exclusivement consacré à résoudre la version en 1D du problème. Nous n'avons pas inclu l'approche proposée par [42] car son adaptation en 1D n'est pas triviale.

3.7 Conclusion

Dans ce chapitre, nous nous sommes intéressés au problème du chevauchement des nœuds sur un dessin de graphe en 1D. Cette problématique a été mise en évidence lors de la conception d'EpidVis et nous avons souhaité la généraliser. Nous avons d'abord identifié 4 critères pour la suppression des chevauchements dans un dessin

Références	C1	C2	C3	C4	Complexité en temps
Misue *et al.* [70]	N	Y	Y	N	$O(\|V\|^2)$
Hayashi *et al.* [47]	N	Y	Y	N	$O(\|V\|^2)$
Marriott *et al.* [68]	N	Y	N	N	$\geq O(\|V\|^2 \cdot log(\|V\|))$
Li *et al.* [64]	N	Y	Y	N	$O(\|V\|^2)$
Dwyer *et al.* [32]	N	Y	N	N	$O(\|V\|log\|V\|)$ and $O((\|V\| \cdot \|C\|)log\|C\|)$
Huang *et al.* [52]	N	Y	Y	N	$O(\|V\|^2)$
Abe *et al.* [1]	N	Y	Y	N	$O(\|V\|^2)$
Notre approche	Y	Y	Y	Y	$O(\|V\|log(\|V\|))$

TABLE 3.1 – Comparaison des différentes approches de suppression de chevauchements de nœuds dans un dessin en 1D : "N" signifie "non", "Y" signifie "oui".

en 1D. Nous avons ensuite proposé un algorithme avec une complexité meilleure que les approches précédentes. Il permet de satisfaire les 4 critères. Nous avons proposé deux études de cas : la première illustre notre approche sur un diagramme en arc, la seconde illustre comment un diagramme en arc généré avec notre approche peut être combiné avec des graphiques silhouette pour la visualisation d'un attribut temporel associé aux nœuds du graphe.

Notre algorithme a été intégré à EPIDVIS et permet de garantir qu'il n'y a pas de chevauchement de mots-clés, notamment après avoir la fusion de nœuds. La Figure 3.14.b illustre l'application de cet algorithme à partir du positionnement de la Figure 3.14.a.

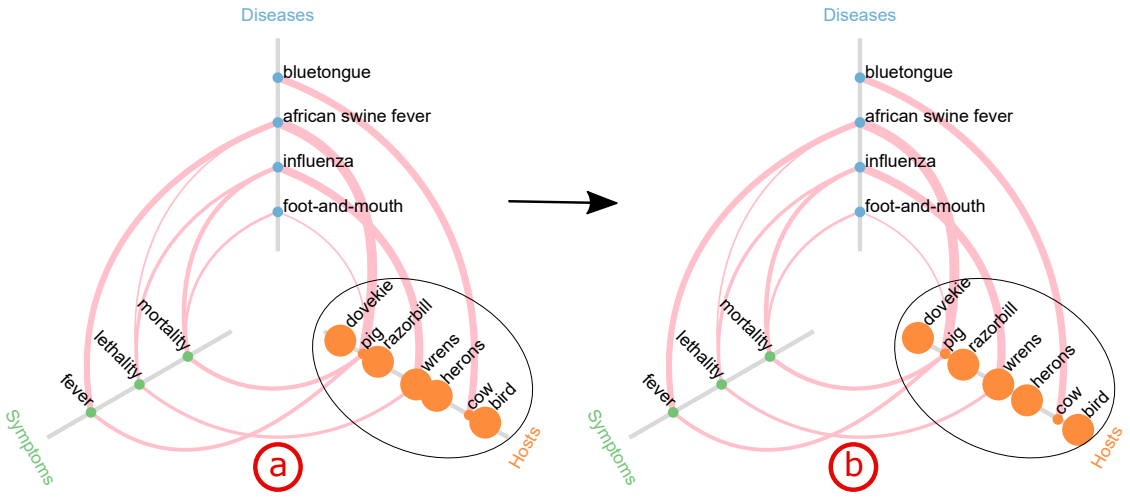


FIGURE 3.14 – Suppression des chevauchements sur l'axe des espèces dans EPIDVIS.

Chapitre 4

Système de visualisation analytique pour la veille épidémiologique

Sommaire

4.1 Introduction

Dans le chapitre 2 nous nous sommes intéressés à la recherche d'informations spécifiques (construction interactive de requêtes pour la recherche d'articles en ligne). L'objectif de ce chapitre est de se focaliser sur les résultats obtenus, et plus particulièrement de proposer aux épidémiologistes une approche leur permettant d'explorer plus facilement une grande quantité d'informations. Comme nous l'avons vu précédemment, les experts utilisent régulièrement des données issues de *sources officielles* internationalement reconnues, comme celles fournies par l'Organisation mondiale de la santé animale (OIE) et le Système de notification des maladies animales (ADNS). Par exemple, l'autorité européenne de sécurité des aliments (EFSA) a utilisé les notifications ADNS pour mettre en évidence la façon dont le virus *African swine fever* initialement signalé dans les pays baltes en 2014 s'est répandu dans la partie orientale de la Pologne, causant ainsi la mort d'*ours* sauvages en grand nombre [27]. Les données provenant des sources officielles sont directement disponibles dans un format bien défini, i.e. avec un étiquetage approprié des indices épidémiologiques concernant les maladies, les espèces, les dates et les localisations.

Dans le chapitre 2, nous avons également mis en évidence qu'il devenait de plus en plus indispensable de parcourir les dépêches de presse pour pouvoir détecter des signaux annonciateurs d'épidémies. Récemment de nouvelles plateformes de surveillance automatique de ces *sources non officielles* ont été développées pour répondre à cet usage [21]. Cependant les données extraites ne sont pas structurées et les différents indices épidémiologiques doivent être extraits soit manuellement, soit à l'aide de techniques de fouille de données [6].

Il est important de noter que dans un contexte de veille épidémiologique, les sources officielles (validées par des experts) et non officielles (issues d'articles de presse) sont très importantes pour les experts.

L'analyse de ces données hétérogènes complexes sans l'aide d'un outil dédié est une tâche ardue. Dans notre contexte, les techniques de visualisation analytique [25] semblent particulièrement bien adaptées pour permettre d'explorer efficacement les différents types de données disponibles. Malheureusement, comme nous le verrons dans la suite de ce chapitre, les outils existants (e.g. [23, 39]) ne sont pas ou peu adaptés à certains besoins exprimés par les épidémiologistes.

Dans ce chapitre, nous présentons EpidNews [44], un nouvel outil de visualisation analytique pour les données épidémiologiques. Il fournit plusieurs vues pour : (1) gérer différentes sources d'information (officielles/non officielles) et différents types de données, (2) visualiser des données géographiques et temporelles et (3) observer les données à différents niveaux d'agrégation.

Le chapitre est organisé de la manière suivante. Dans un premier temps, en section 4.2, nous présentons la problématique qui nous permet de déterminer une liste de critères que doit satisfaire une application de visualisation d'épidémies. Les principales approches existantes sont, par la suite, étudiées pour nous permettre de faire un premier bilan et de mettre en évidence les approches qui peuvent répondre aux

critères spécifiés. Dans la section 4.3, nous présentons les différentes visualisations proposées ainsi que leurs modes d'interaction. Une étude de cas menée par des épidémiologistes avec EpidNews est proposée dans la section 4.4. Comme précédemment, le chapitre se termine par une conclusion dans la section 4.5.

4.2 Problématique et état de l'art

Pour développer une plateforme qui réponde aux besoins des épidémiologistes, nous avons mis en évidence à travers différentes réunions, les principales questions qui sont primordiales pour les utilisateurs.

- Est-il possible de connaître le lieu d'apparition d'une maladie ("où") ?
- Comment savoir le moment d'apparition d'une maladie ("quand") ?
- Quelles sont les informations disponibles au sujet de maladies ou des espèces dans un lieu précis et à une date précise ("quoi") ?
- Est-il est possible de récupérer des informations à partir de sources multiples et d'identifier ces sources ?
- Sachant que les maladies, les symptômes et les espèces sont fortement liés, est-il possible d'observer ces trois catégories et présenter plusieurs données de chaque catégorie en même temps ?
- Est-il possible de résumer les informations ?
- Est-il possible de se focaliser sur une région spécifique ou sur une période de temps choisie ?

L'intérêt de ces questions est de nous avoir permis d'identifier, avec les utilisateurs, les principaux critères que doit vérifier une application de visualisation d'épidémies et que nous décrivons dans la section suivante.

4.2.1 Les critères

Nous distinguons deux groupes de critères principaux. Le premier concerne ce que nous appelons les critères de base. Ils rassemblent les attributs fondamentaux liés notamment aux informations spatio-temporelles. Le second groupe correspond à des critères transversaux. Il s'agit de fonctionnalités supplémentaires qui peuvent être ajoutées aux critères de base afin de faciliter l'exploration des données.

4.2.1.1 Les critères de base

Les données spatio-temporelles permettent de représenter des informations aussi bien dans l'espace que dans le temps. Elles prennent donc en charge différents attributs et la manière dont ils sont corrélés. Dans [75], Peuquet a proposé un modèle triadique (Cf. Figure 4.1) pour décrire comment ces attributs sont reliés et peuvent

être utilisés à travers trois composantes : le *où* (la localisation), le *quand* (le temps) et le *quoi* (l'information). L'auteur met particulièrement en avant les corrélations suivantes :

- **où** + **quand** → **quoi** : permet de décrire les informations (quoi) qui sont présentes à des emplacements (où) et à un moment donné (quand) ;
- **où** + **quoi** → **quand** : permet de décrire le moment ou la période (quand) durant laquelle des informations (quoi) sont localisées dans un lieu donné (où) ;
- **quoi** + **quand** → **où** : permet de décrire l'emplacement (où) dans lequel des informations (quoi) sont présentes à un moment donné (quand).

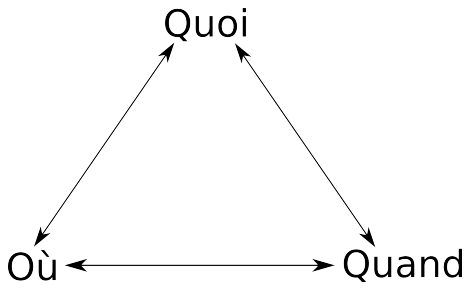


FIGURE 4.1 – Composants de base du modèle triadique.

Comme nous pouvons le constater, ce modèle triadique est particulièrement pertinent dans notre contexte. Par exemple, à l'aide du *où* et du *quoi*, il est possible de connaître à partir d'un lieu et d'une maladie précise le moment durant lequel a débuté une épidémie. Dans les sections suivantes, nous précisons plus en détail les représentations visuelles associées à ces composantes.

C1 : Où (visualisation géographique)

La visualisation géographique est la représentation des informations (**quoi**) qui sont associées à des espaces géographiques. Traditionnellement la représentation la plus adaptée est la carte. Un exemple frappant lié au domaine qui nous intéresse ici a été proposé par John Snow, médecin britannique au 19ème siècle, pour mettre en évidence le foyer d'une épidémie de choléra à Londres. Pour ce faire, il a reporté sur une carte le nombre de décès à l'aide de barres noires perpendiculaires aux rues (Cf. vue générale Figure 4.2). En observant cette carte, on peut remarquer que la densité des morts varie selon les lieux et qu'elle se concentre autour d'un puits (Cf. vue détaillée Figure 4.2). Il se trouve que ce puits était justement celui qui était contaminé. Il existe actuellement de très nombreuses plateformes logicielles disponibles pour visualiser de l'information sur des cartes (e.g. Google Maps [1], Bing Maps [2], Open Street Maps [3], etc.).

1. https://maps.google.fr
2. www.bing.com/maps
3. www.openstreetmap.org

FIGURE 4.2 – Vue générale d'une carte de décès dus à une épidémie de choléra à Londres en 1854 et vue détaillée de la région de *Broad Street* infectée par le choléra.

C2 : Quand (visualisation du temps)

Ce critère correspond à la représentation des informations au cours du temps. Il est particulièrement utile pour observer, par exemple, l'évolution d'une information à différents pas de temps. Il existe dans la littérature de très nombreuses approches pour visualiser les informations temporelles. Dans [3], Aigner et al. proposent un état de l'art sur le sujet. Ils ont aussi créé un site spécifique [4] qui offre la possibilité d'accéder à de très nombreuses visualisations (Cf. Figure 4.3). Elles sont présentées sous une forme matricielle et il est possible de sélectionner celles qui sont les plus appropriées à l'aide d'un mécanisme de filtres. En particulier, l'utilisateur peut sélectionner le type de données (univariées ou multivariées, spatiales ou abstraites, etc.), la manière de représenter le temps (linéaire, cyclique, par périodes, par intervalles, etc.), le type de visualisation (statique ou dynamique, 2 ou 3 dimensions, etc).

EventRiver [67] est un exemple d'approche basée sur la représentation d'intervalles de temps. Des dépêches d'actualités portant sur le même sujet sont regroupées et représentées sous la forme de bulles. Chaque bulle a une couleur spécifique et sa taille représente le nombre de dépêches. Les groupes sont ordonnés par ordre d'apparition. La Figure 4.4 illustre différents événements intervenus au cours du temps. Cette visualisation permet de mettre en évidence les évolutions des sujets d'actualité. On peut par exemple observer de nombreuses dépêches dans le groupe rouge du 1er au 15 août et beaucoup moins après.

Le système TimeLineCurator [41] [5], non mentionné dans timeviz.net, est une autre approche de visualisation d'événements au cours du temps. Il offre la possibilité, après avoir récupéré des informations temporelles dans des documents textuels, de positionner un événement sur une ligne temporelle. Le temps est représenté par un

4. http://timeviz.net/
5. http://timelinecurator.org

FIGURE 4.3 – Un aperçu du site Timeviz.net.

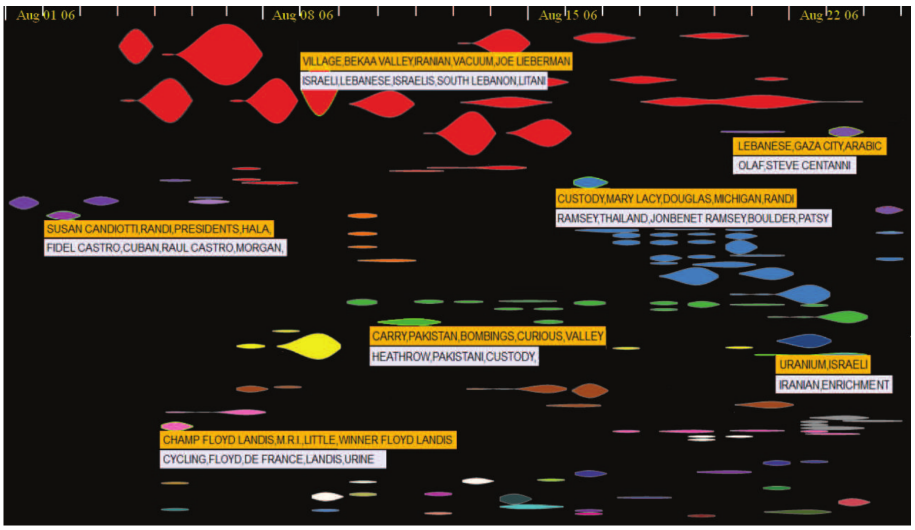


FIGURE 4.4 – Actualités de CNN du 1er au 22 août (29,211 documents sous-titrés) dans EventRiver.

axe horizontal. Les événements sont positionnés sous la forme de cercles s'il s'agit de dates précises ou bien sous la forme de barres temporelles s'il s'agit de périodes. Différentes couleurs sont utilisées pour mettre en évidence des sujets particuliers. Par exemple, la Figure 4.5 illustre les événements liés à la musique Pop entre 1956 et 2015. Les couleurs représentent les différents pays. L'événement sélectionné dans la figure correspond au Progg (mouvement musical Suédois) entre 1960 et 1980.

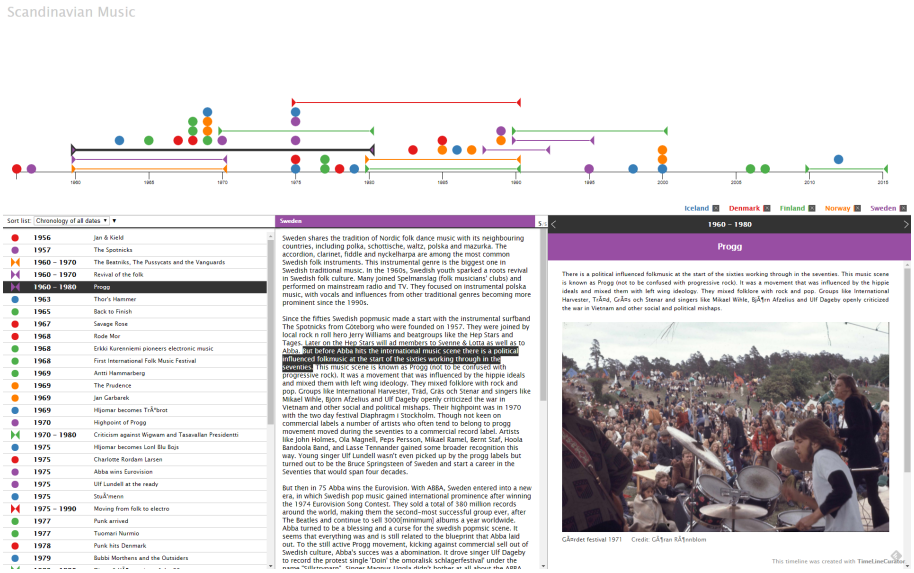


FIGURE 4.5 – Un exemple de TimeLineCurator avec la musique Pop.

C3 : Quoi (visualisation des sujets d'intérêt)

Les sujets d'intérêts (**quoi**) peuvent être visualisés selon un contexte spatio-temporel (**où, quand**). Ces sujets peuvent être associés à des informations spatio-temporelles plus précises notamment via des interactions que nous préciserons ultérieurement. Par exemple, la Figure 4.6 représente une carte de maladies au Brésil avec une information complémentaire sur la grippe tropicale (**quoi**) à Acrelândia - Brésil (**où**) le 12 janvier 2018 (**quand**).

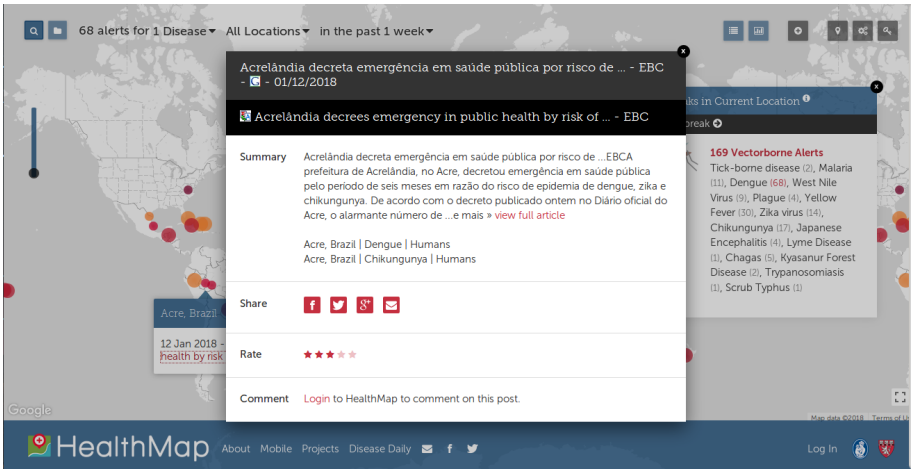


FIGURE 4.6 – Un exemple d'article sur la grippe tropicale à Acrelândia au Brésil le 12 janvier 2018.

Les informations (**quoi**) peuvent correspondre à différents types de données (e.g. [62]). Certaines approches permettent visuellement de faire la distinction entre les

types de manière à en faciliter l'analyse. Par exemple, la carte de la Figure 4.7 illustre différents types de données : établissements de sports, d'éducation, etc.

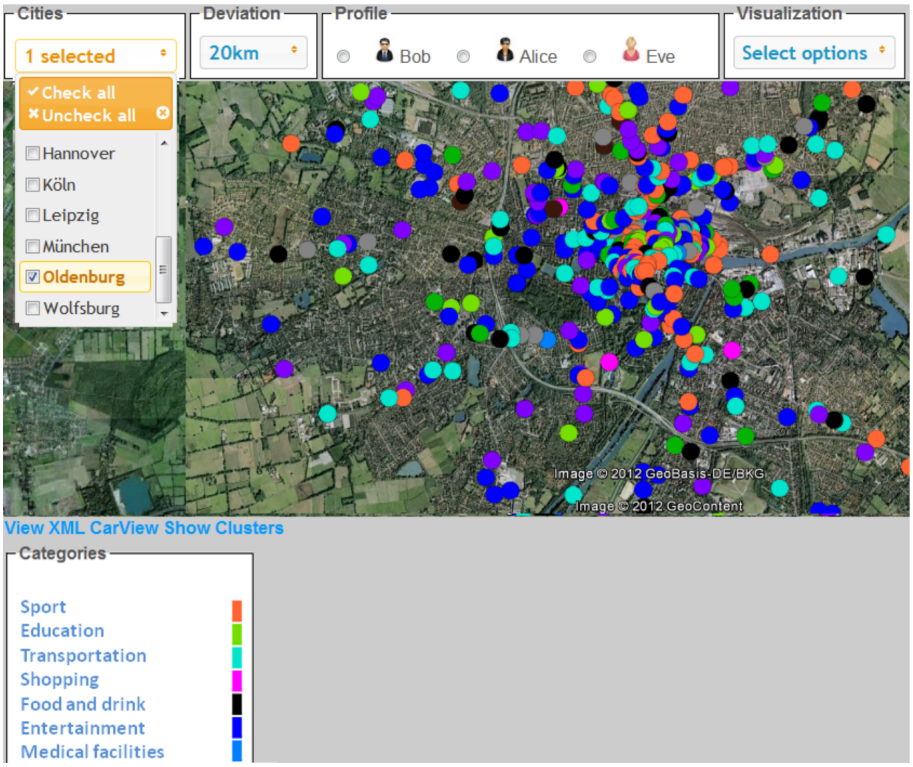


FIGURE 4.7 – Visualisation de différents types de données via des cercles de différentes couleurs.

4.2.1.2 Les critères transversaux

Pour pouvoir répondre aux différentes questions des utilisateurs, les critères de base peuvent se révéler suffisants. Nous proposons de les compléter par des critères transversaux que nous détaillons dans les sections suivantes.

C4 : Résumer l'information

Visualiser un résumé ou une agrégation d'informations peut être utile à l'utilisateur afin de lui donner une vue d'ensemble des données. Nous présentons tout d'abord les approches qui s'intéressent uniquement à la mise en évidence d'informations agrégées sur une carte, i.e. le critère **où**.

Il existe généralement deux types d'approches. Les premières sont basées sur des régions aux contours spécifiques : ce sont principalement les cartogrammes et les cartes choroplèthes. Au contraire, les secondes ne sont pas basées sur des régions bien délimitées. Elles permettent de représenter une densité d'informations, par exemple sous la forme de cartes de chaleur. La manière de présenter les informations peut,

dans ce cas, se faire en faisant varier le rayon d'un cercle sur une carte ou en jouant sur les gradients de couleurs.

La Figure 4.8 illustre les trois types de cartes. Le cartogramme [6] de la Figure 4.8.a montre la population par pays. La carte choroplèthe de la Figure 4.8.b [29] représente les prix du gaz aux États-Unis. Enfin la carte de chaleur de la Figure 4.8.c [73] montre les positions des utilisateurs de Brightkite entre avril 2008 et octobre 2010.

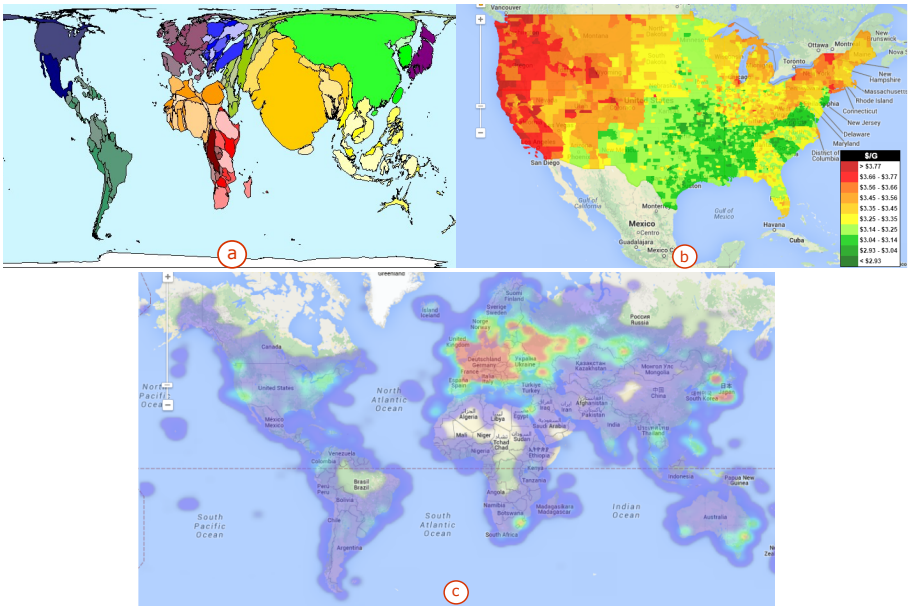


FIGURE 4.8 – Exemple de résumés d'informations liés à la localisation : cartogramme **(a)**, carte choroplèthe **(b)** et carte de chaleur **(c)**.

VoroGraph [31] est une approche qui permet de prendre en compte aussi bien les critères **où** et **quand** pour résumer l'information. Il s'agit d'une plateforme de simulation de propagation de maladies. Les informations peuvent être visualisées tout d'abord sous la forme d'une carte choroplèthe mettant en évidence des régions proches possédant de fortes interactions (Cf. Figure 4.9.a). De plus, en se basant sur une tesselation de Voronoi, la carte est transformée de manière à rapprocher les zones à forte interactions tout en maintenant l'aspect initial de la carte (Cf. Figure 4.9.b). Enfin, une représentation sous la forme d'un graphe où les nœuds correspondent à des régions issues de la tesselation précédente et les liens les corrélations entre les régions est intégrée. Ces liens, qui peuvent être pondérés, permettent de mettre en évidence les corrélations entre des régions même très éloignées. La Figure 4.9 illustre la visualisation de la propagation de la maladie *influenza* dans le monde à un instant donné. Les trois types de visualisation sont présentes : la carte, la tesselation de Voronoi et le graphe associé.

Visuellement le temps, peut être résumé de multiples façons : cercles, diagrammes, matrices. Ces dernières peuvent et sont souvent basées sur l'agrégation des heures en

6. www.worldmapper.org

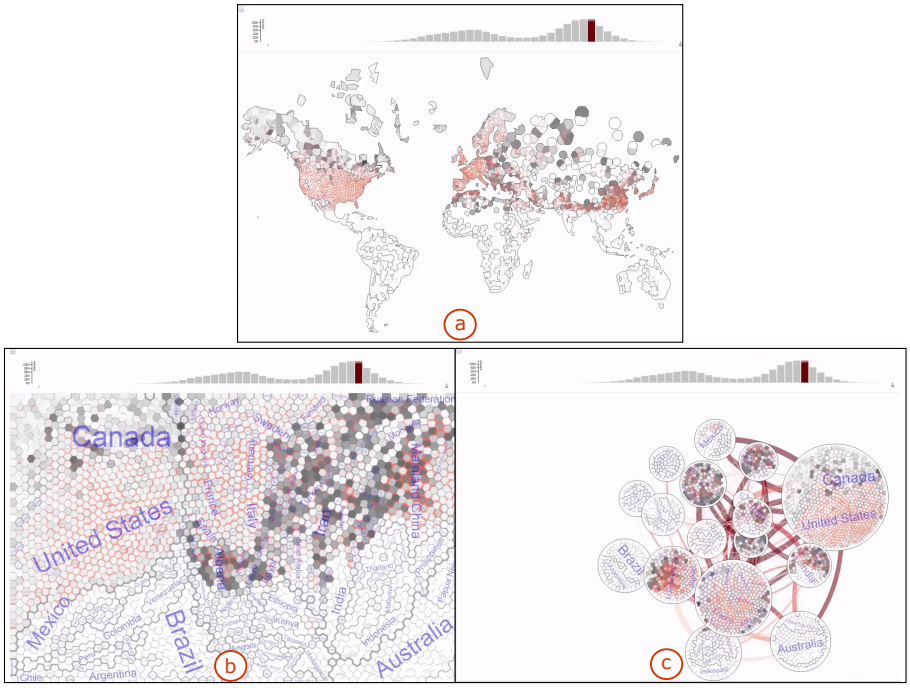


FIGURE 4.9 – Résumé d'informations et d'espace dans VoroGraph. Résumé d'espaces géographiques via une carte **(a)**, à l'aide d'une tesselation de Voronoi **(b)**, et sous la forme d'un graphe **(c)**.

jours, des jours en semaines, des semaines en mois, etc. Les informations associées sont donc agrégées de façon similaire : cumul de valeurs par exemple. Dans [83], les auteurs proposent une approche matricielle qui permet d'agréger des informations géolocalisées à différents niveaux de granularité temporelle. Ils utilisent pour cela une matrice de pixels (Cf. Figure 4.10) pour représenter l'information spatiale de la donnée à un instant précis. La matrice est divisée en trois parties : données quotidiennes, données mensuelles et données annuelles (Cf. Figure 4.10 de haut en bas). Chaque colonne de la matrice représente une année de données. Les cellules de la région la plus haute représentent les mois. Ils montrent tous les détails en codant, par couleur, les pixels individuels dans une cellule en fonction des valeurs quotidiennes. La région du milieu montre des données agrégées. Ici, les cellules ne sont plus subdivisées en pixels, elles sont colorées uniformément. La couleur représente l'agrégation des valeurs journalières en une seule valeur mensuelle (idem pour la région du bas). Dans la rangée du bas, 12 valeurs mensuelles sont agrégées en une seule valeur pour l'année 1971.

C5 : Recherche interactive

Les interactions correspondent à un dialogue entre l'utilisateur et la visualisation, permettant ainsi à cet utilisateur d'explorer l'ensemble de données afin de découvrir de nouvelles informations [98]. Elles peuvent, bien entendu, être appliquées aux

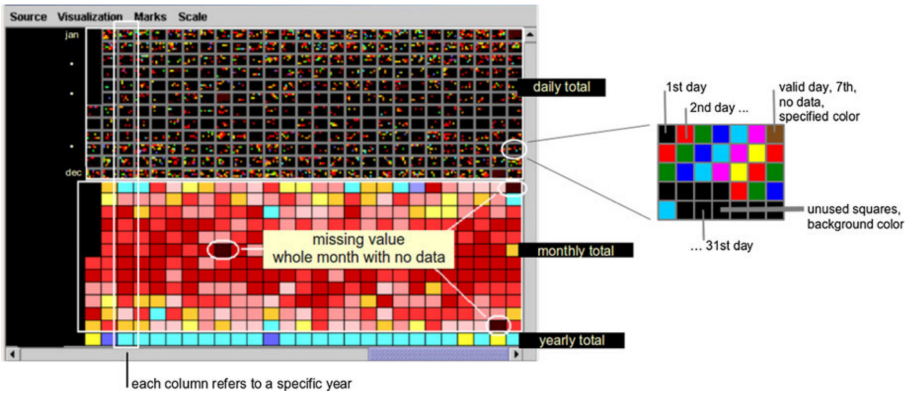


FIGURE 4.10 – Représentation matricielle du temps. Application à des données de précipitation au Brésil au cours de l'année 1971 avec différentes granularités temporelles.

trois critères de base et offrent ainsi des facilités de filtrage d'informations. Par exemple, l'utilisateur peut vouloir connaître des localisations précises à une période donnée. Elles peuvent également être utilisées pour faire des comparaisons. Par exemple, dans [37], Ferreira et al. proposent une visualisation interactive qui permet de créer des requêtes selon différents critères de base. La Figure 4.11 représente une comparaison de trajets en taxi de Lower Manhattan aux aéroports JFK (John Fitzgerald Kennedy de New York) et LGA (La Guardia de New York) en mai 2011. La requête de gauche sélectionne les voyages qui ont eu lieu le dimanche, tandis que celle de droite sélectionne les voyages qui ont eu lieu les lundis. Les utilisateurs spécifient ces requêtes en sélectionnant visuellement des régions sur la carte et en les connectant. En plus, ils peuvent sélectionner les résultats représentés sur la carte. Des nuages de points en bas des cartes montrent les relations entre les heures et les durées des trajets. Les points sont colorés en fonction de la contrainte spatiale représentée par les flèches entre les régions : les déplacements vers JFK en bleu et les déplacements vers le LGA en rouge. On peut observer que la plupart des voyages effectués le lundi entre 15h et 17h prennent beaucoup plus de temps que les voyages le dimanche.

C6: Présentations des sources

Les informations peuvent être issues de plusieurs types de sources (Twitter, moteurs de recherche, données officielles, etc). L'objectif de ce critère est de savoir s'il est possible d'identifier visuellement ces types au sein d'une carte.

Les approches traditionnelles utilisent des glyphes ou des couleurs différentes pour mettre en évidence des types de sources multiples. Par exemple, la Figure 4.12 montre Cluster Bulls-Eye [85, 86], une visualisation qui offre aux utilisateurs des moyens d'observer les intersections de groupes de documents, chaque groupe étant composé de 100 documents récupérés via un moteur de recherche différent : Google,

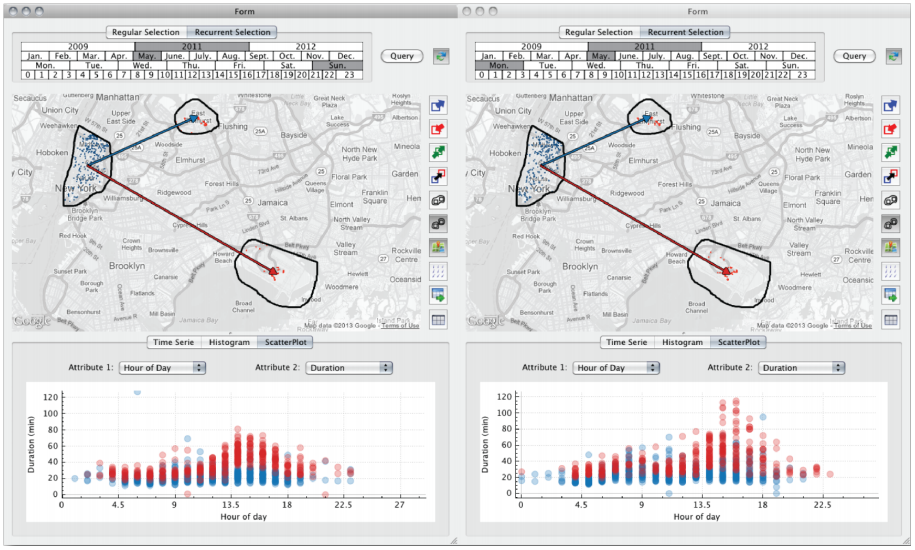


FIGURE 4.11 – Comparaison de trajets en taxi de Lower Manhattan aux aéroports JFK et LGA en mai 2011.

Teoma, AltaVista, Lycos et MSN. Cette visualisation utilise des couleurs et une organisation circulaire des groupes afin d'identifier les types de sources.

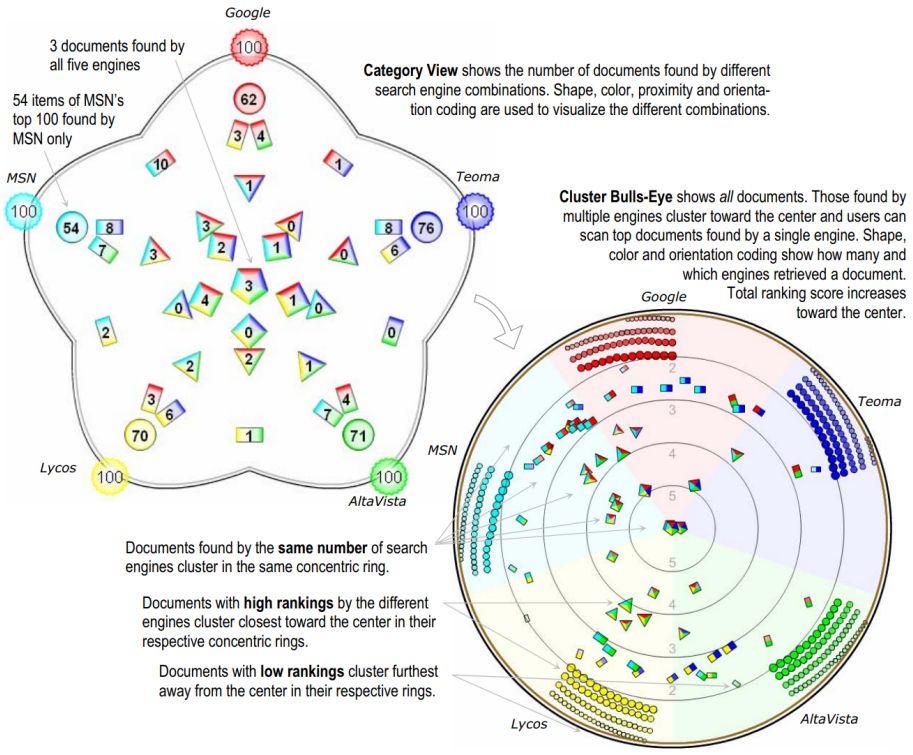


FIGURE 4.12 – Cluster Bulls-Eyes : visualisation des premiers documents récupérés via différents moteurs de recherche.

Après avoir présenté les différents critères, la section suivante détaille la manière dont ils sont pris en compte dans les principales plateformes de veille épidémiologique.

4.2.2 Plateformes de visualisation d'épidémies

Les plateformes de visualisation d'épidémies peuvent être très diverses en fonction des objectifs visés. Il existe, par exemple, des approches spécifiques basées sur des simulations qui permettent d'étudier les propagations des épidémies en fonction de différents critères [31, 92], ou bien dans des données biologiques [66]. Cependant, la plupart des approches satisfont les critères de base, i.e. extraire visuellement sur différentes périodes de temps ou de lieux un ensemble d'informations. La différence essentielle est liée aux critères transversaux qui sont souvent spécifiques aux applications visées. Dans cette section, nous nous intéressons principalement aux systèmes de surveillance d'épidémies qui sont proches de notre contexte. Ces derniers permettent d'afficher des alertes qui peuvent être officielles ou non officielles. Ils permettent également de suivre les évolutions des différentes maladies au cours du temps.

Empres-i

Empres-i[7] [24] est une application qui utilise une carte interactive (C1) pour représenter des informations associées à des maladies (Cf. Figure 4.13). Les données sont issues de sources officielles (e.g. FAO - Food and Agriculture Organization of the United Nations, laboratoires de référence, autorités nationales) et non officielles (articles d'actualités).

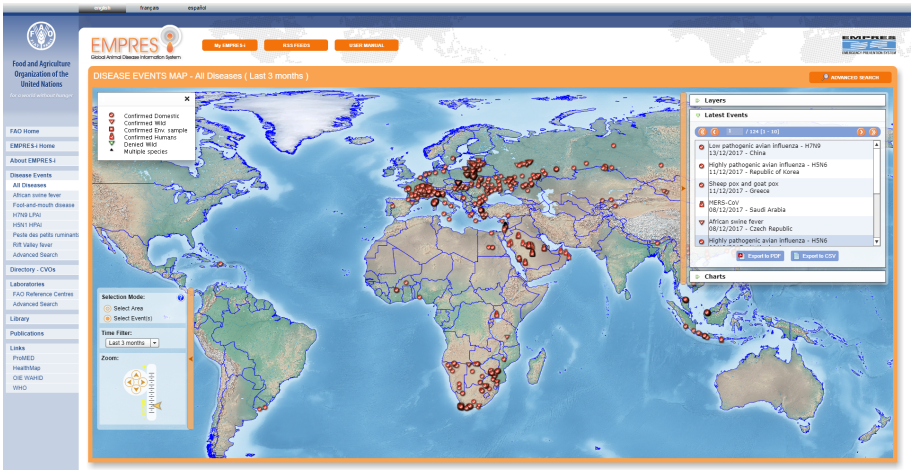


FIGURE 4.13 – Empres-i.

7. http://empres-i.fao.org/eipws3g/

Des interactions permettent d'obtenir des informations complémentaires associées aux maladies (C3, C4) : sources (C6), espèces, statut (confirmé ou non), date, type de résultats (positif ou non), etc. La Figure 4.14 montre un exemple avec la *peste porcine Africaine*.

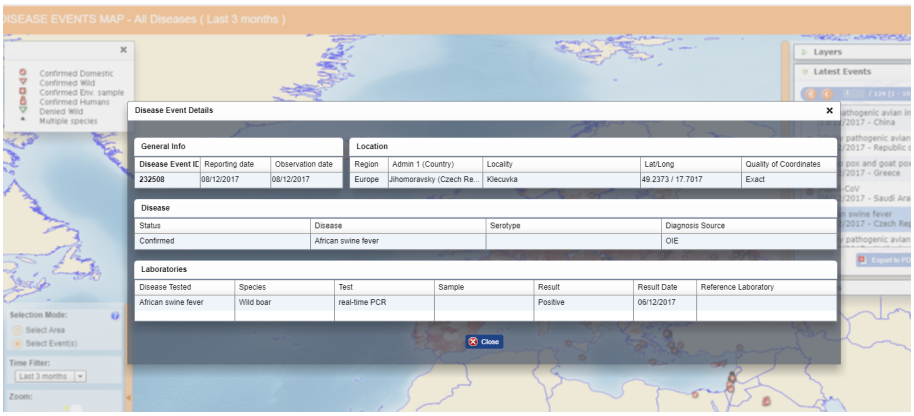


FIGURE 4.14 – Résumé d'informations sur la *peste porcine Africaine*.

L'utilisation de glyphes sur la carte permet de représenter les maladies. Ils peuvent être différents en fonction de l'information qu'ils représentent. Par exemple, dans la Figure 4.15, si une maladie est confirmée comme infectant des animaux domestiques, elle est représentée par un cercle alors qu'un triangle est utilisé pour les maladies confirmées infectant des animaux sauvages. De très nombreuses interactions sont proposées (C5). Comme nous pouvons le constater sur la Figure 4.15, l'utilisateur peut, à l'aide de la boîte à outils située à gauche de la carte, sélectionner des périodes de temps (C2), des emplacements géographique en créant un rectangle sur la carte et en activant le bouton "Select Area", sauvegarder et sélectionner des événements, etc.

HealthMap

HealthMap [8] est également une approche de surveillance d'épidémies qui utilise une carte interactive (C1). Les données sont obtenues à partir de données officielles (rapports de surveillance, alertes officielles validées) ou non officielles (fil d'actualités, comptes d'experts (e.g. ProMED-mail)). Ces dernières sont traitées à l'aide de méthodes de fouille de textes afin d'en extraire les informations utiles par exemple pour le positionnement sur la carte. La Figure 4.16 montre un exemple de l'application.

De nombreuses techniques permettant de résumer l'information sont disponibles dans le système (C4). Les maladies sont représentées à l'aide des cercles. Le nombre de maladies pour une localisation donnée est mis en évidence à l'aide de la couleur en utilisant un gradient variant du jaune (valeur minimale) au rouge (valeur maximale). De plus le rayon du cercle est utilisé pour mettre en évidence les maladies spécifiques

8. http://healthmap.org

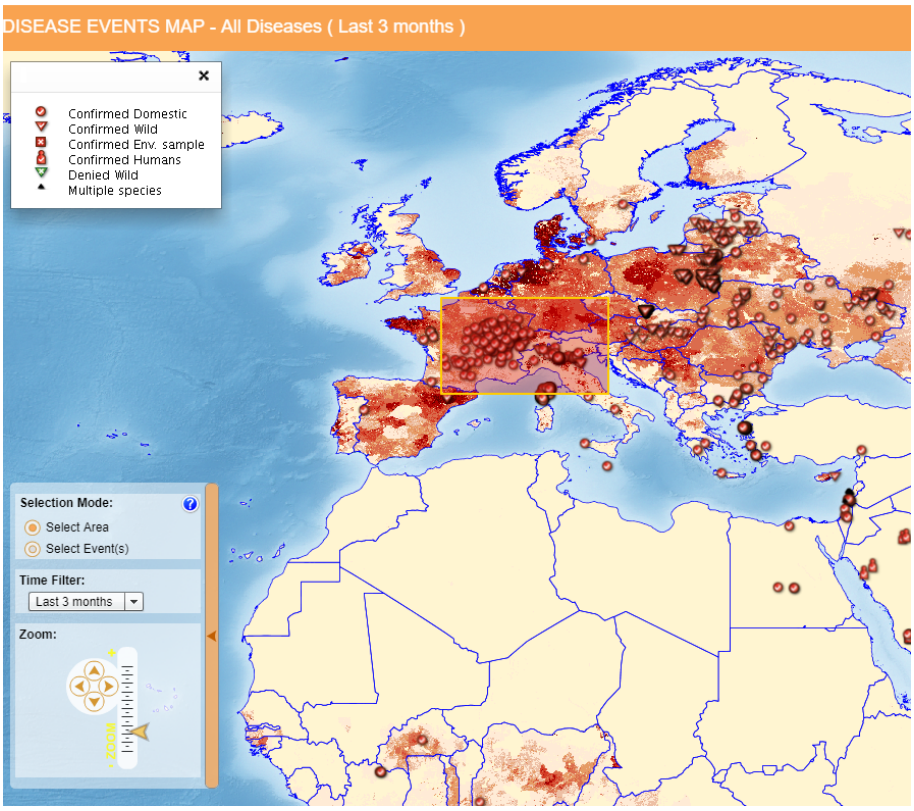


FIGURE 4.15 – Interactions d'Empres-i : sélection de données dans une région européenne.

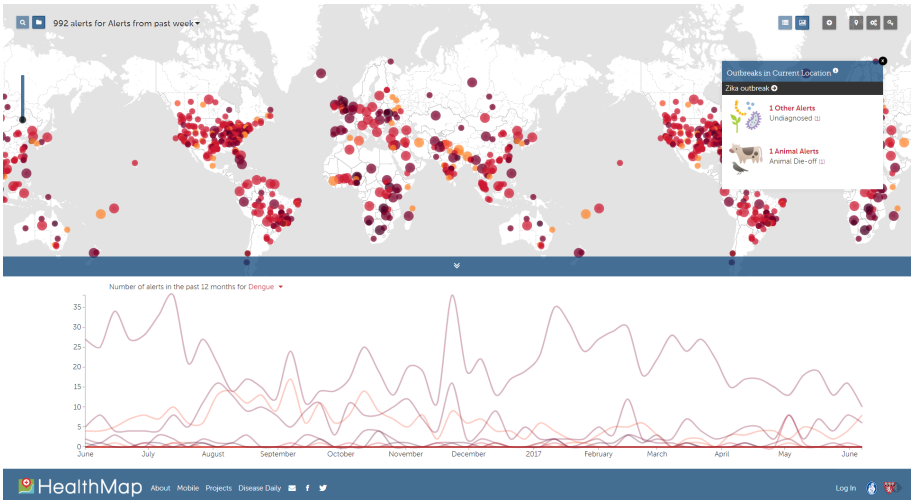


FIGURE 4.16 – HealthMap.

à un pays (rayon important) ou à une région (rayon faible). La Figure 4.17 montre les différents niveaux de granularité. Le nombre d'alertes par estampille temporelle est également proposé comme le montre le bas de la Figure 4.16.

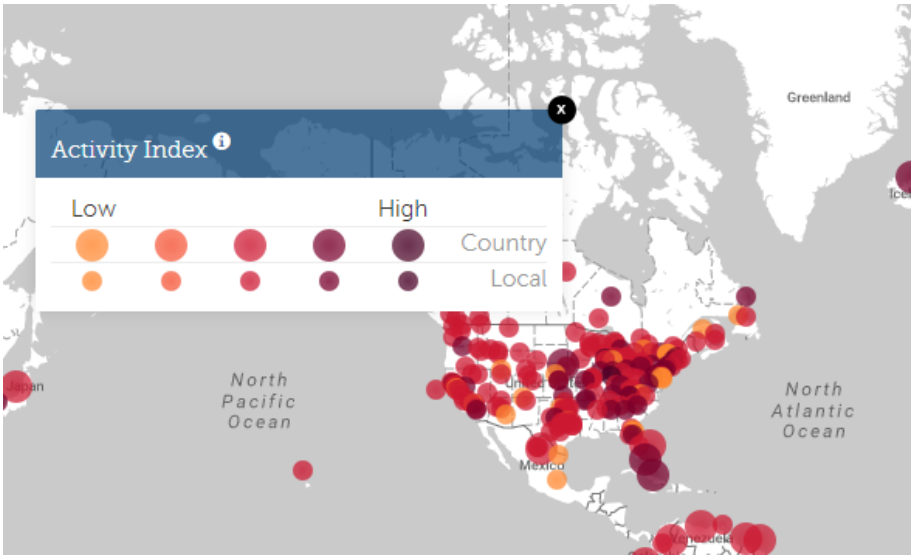


FIGURE 4.17 – Variation du rayon en fonction de l'index d'activité.

La distribution géographique des espèces (C4) est représentée sous la forme d'une carte de chaleur. Par exemple, la Figure 4.18 montre la distribution géographique de l'espèce *Albopictus* représentée via une carte de chaleur et de la maladie *Zika* représentée à l'aide de cercles le 31 mars 2017.

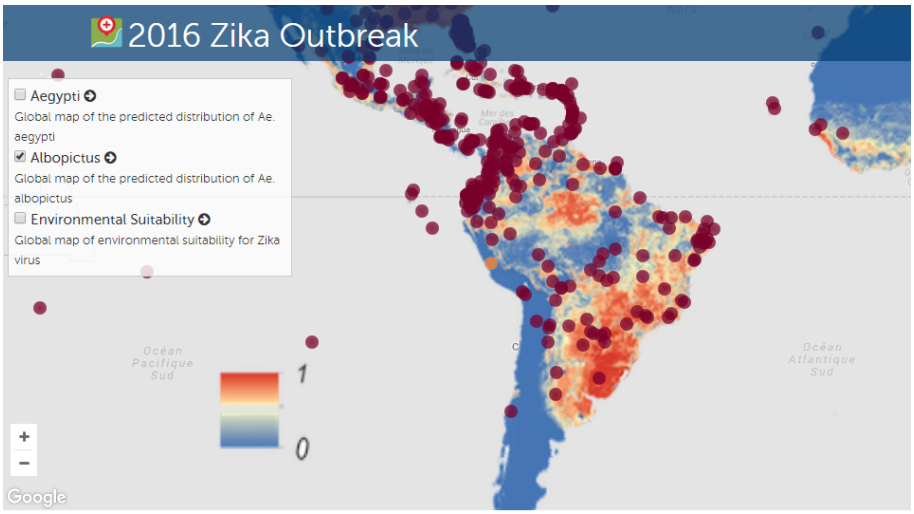


FIGURE 4.18 – La distribution de l'espèce Albopictus (carte de chaleur) et de la maladie Zika (cercles) le 31 mars 2017.

HeathMap offre de nombreux mécanismes d'interaction (C5). Comme le montre la Figure 4.19, l'utilisateur peut filtrer l'emplacement, le temps (C2), les maladies ou les espèces (C3).

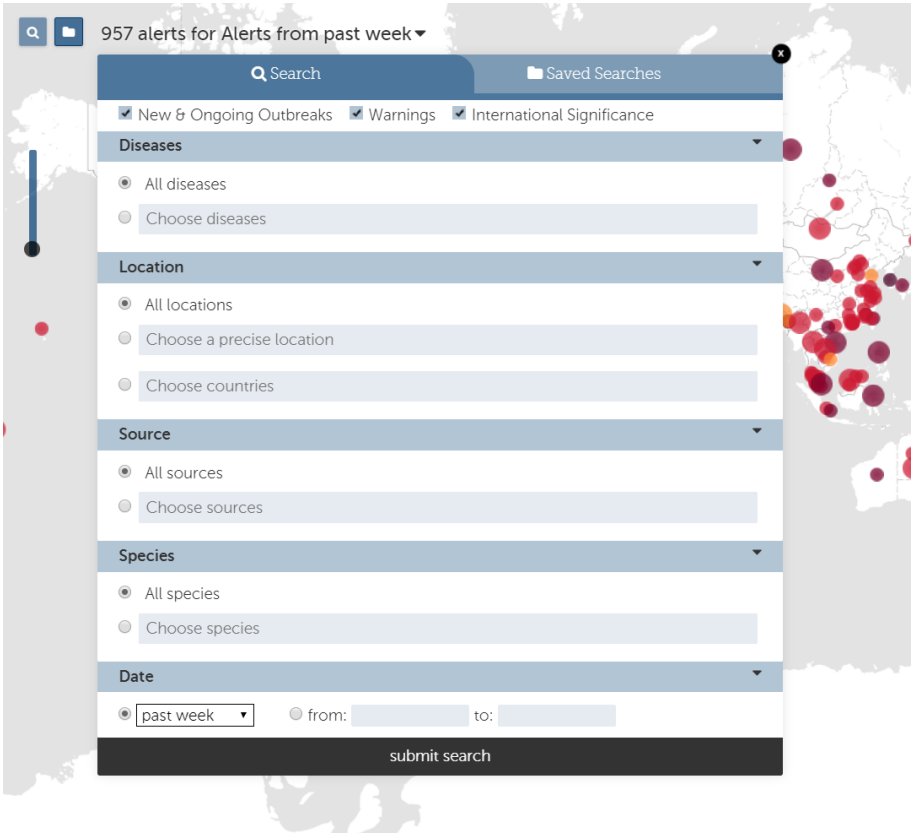


FIGURE 4.19 – Les interactions de HeathMap via une boîte à outils.

Les sources sont représentées par des icônes (C6) apparaissant quand l'utilisateur sélectionne un cercle associé à une maladie. Il faut noter que ces icônes sont uniques pour chaque source. Par exemple, l'utilisateur peut choisir des sources qui ont, au préalable, été sauvegardées dans le système (Cf. la Figure 4.20.a) ou bien des sources actuelles. La Figure 4.20.b montre la maladie *hantavirus* au Chili le 20 janvier 2018. Cette information est récupérée à partir de ProMed Mail (l'icône à gauche du texte précise l'origine de la source).

Les deux premières approches que nous venons de décrire sont des outils de surveillance d'épidémies et intègrent notamment des données fréquemment mises à jour. Les deux suivantes possèdent des fonctionnalités assez similaires comme nous allons le voir mais leur objectif est assez différent car elles ne sont pas forcément dédiées aux épidémies et se focalisent plus sur des données agrégées issues souvent d'organismes officiels.

GapMinder

GapMinder [9] est un système de visualisation de données statistiques. Les données utilisées sont issues principalement d'organisations comme l'Organisation Mondiale

9. https://gapminder.org/tools

FIGURE 4.20 – Représentation de sources dans HeathMap : **(a)** représente des sources sauvegardées, **(b)** représente des sources avec les icônes associées.

de la Santé (OMS), l'Organisation des Nations Unies pour l'éducation la science et la culture - UNESCO, etc. Les informations peuvent être affichées sur une carte (C1) et il existe de nombreux moyens pour résumer (C4) ou mettre en évidence les informations disponibles (C3). Ainsi les maladies dans une région sont représentées par des cercles avec des rayons proportionnels au nombre de cas. Les couleurs sont spécifiques à chaque continent. La Figure 4.21 montre une carte de cercles de tailles différentes représentant les adultes qui ont le sida (HIV) entre 15 et 49 ans. Les cercles représentent les régions infectées. Leur rayon est proportionnel au nombre d'infections dans ces régions. Une ligne de temps (C2) est associée à cette carte. L'utilisateur peut également visualiser la distribution des maladies au cours du temps ou sélectionner une année particulière.

De très nombreux modes d'interactions sont proposés. Par exemple, une carte et une liste à droite de la carte centrale permettent aux utilisateurs de sélectionner des continents et des régions. Dans la Figure 4.22 l'utilisateur sélectionne l'Afrique. Par conséquent les données de la carte principale concernant les autres continents sont de couleur moins vive et celles concernant l'Afrique sont mises en surbrillance et étiquetées.

L'utilisateur peut aussi sélectionner (C5) un cercle et consulter la valeur qu'il contient, i.e. le nombre de maladies (C3). Cette valeur est affichée à gauche de la carte. Par exemple, dans la Figure 4.23, le rayon du cercle indique que le nombre de personnes infectées par le HIV au Zimbabwe est élevé.

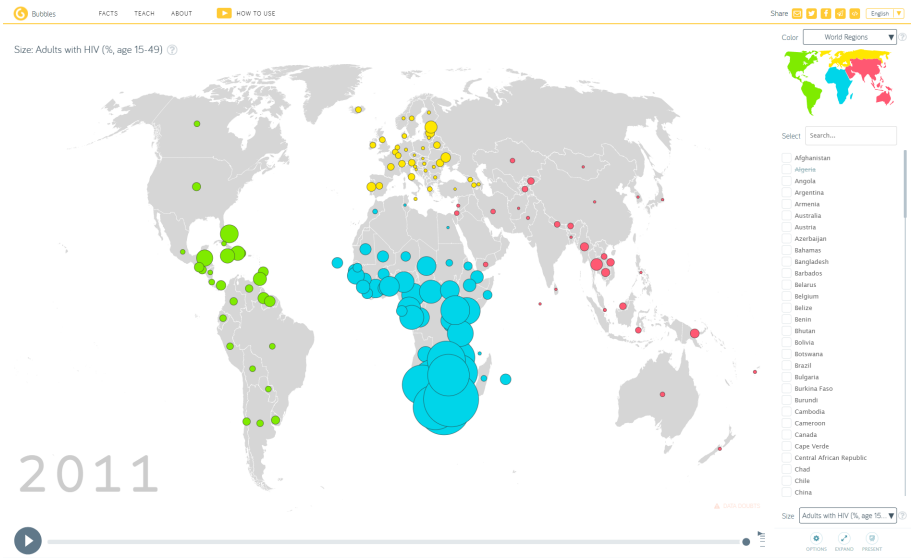


FIGURE 4.21 – GapMinder. Visualisation d'épidémie HIV.

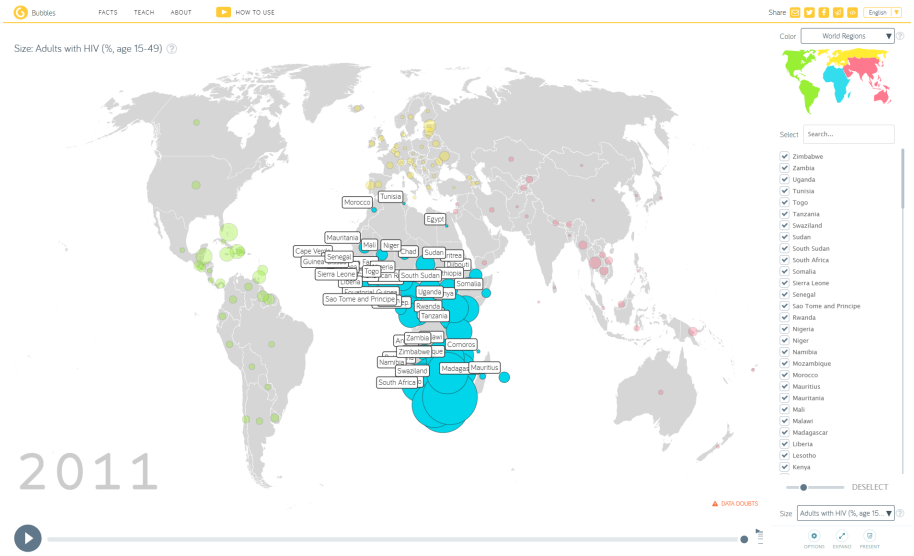


FIGURE 4.22 – HIV en Afrique en 2011.

GapMinder utilise des sources issues de très nombreux domaines. Elles sont présentées (C6) sous la forme d'une liste (dans les boutons en bas et en haut) comme le montre la Figure 4.24.

Epi Visualization

Epi Visualization[10] proposé par l'IHME (Institute for Health Metrics and Evaluation) est un outil interactif pour présenter et explorer des données de santé ou

10. https://vizhub.healthdata.org/epi

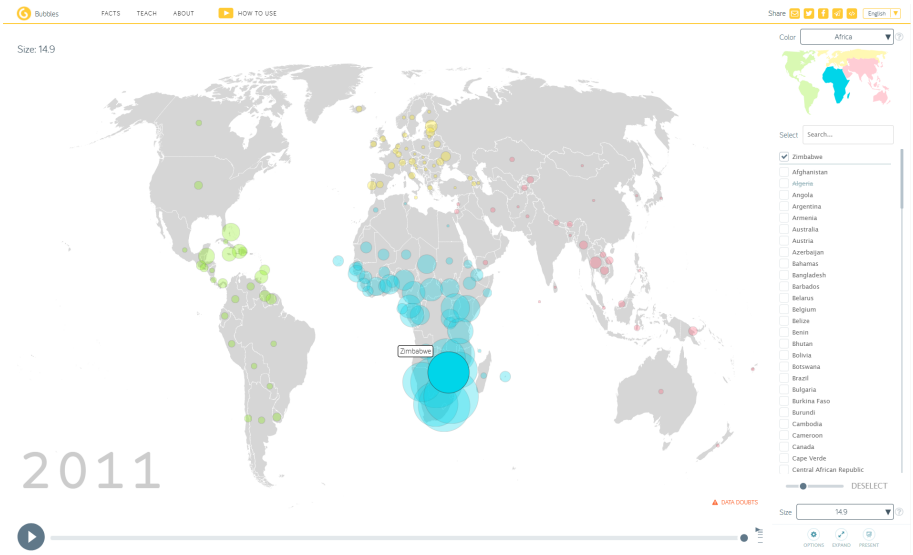


FIGURE 4.23 – HIV au Zimbabwe en 2011.

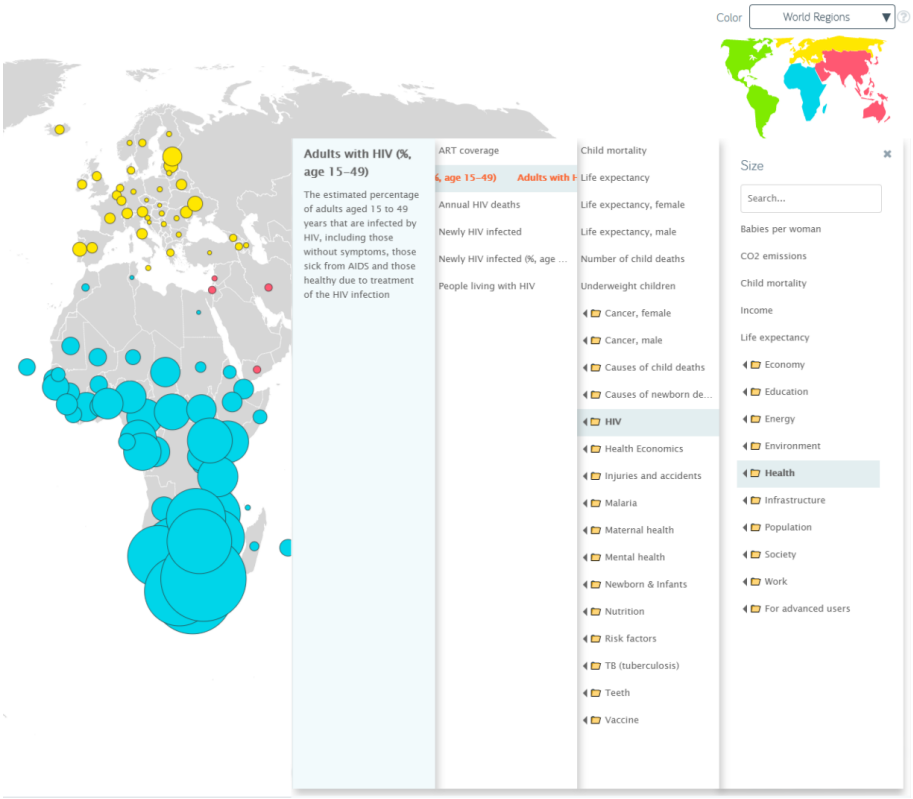


FIGURE 4.24 – Les sources proposées dans GapMinder.

des résultats d'estimations sur des épidémies. Les ressources utilisées sont assez variables : résultats de sondages, recensements, statistiques de l'état-civil et données

relatives à la santé issus d'organismes officiels. Les informations peuvent être visualisées sur une carte (C1). De nombreuses approches de résumés d'information (C4) sont offertes : carte choroplèthe, nuages de points et diagramme en ligne. Des couleurs permettent de mettre en évidence les données en fonction des situations géographiques. Enfin, de nombreuses interactions (C5) sont également proposées. Par exemple l'utilisateur peut sélectionner des zones géographiques, des périodes de temps (C2), des sources (C6), avoir des informations précises (C3) sur de nombreux attributs (age, types de mesures, etc). La Figure 4.25 montre le système avec des données portant sur l'épilepsie dans le monde. L'utilisateur peut sélectionner des pays en utilisant l'échelle affichée en bas de la figure. Cette échelle est représentée par une ligne colorée où les couleurs correspondent aux valeurs associées à l'attribut suivi (ici l'âge normalisé). Les couleurs sont alors reportées sur les différents pays de la carte.

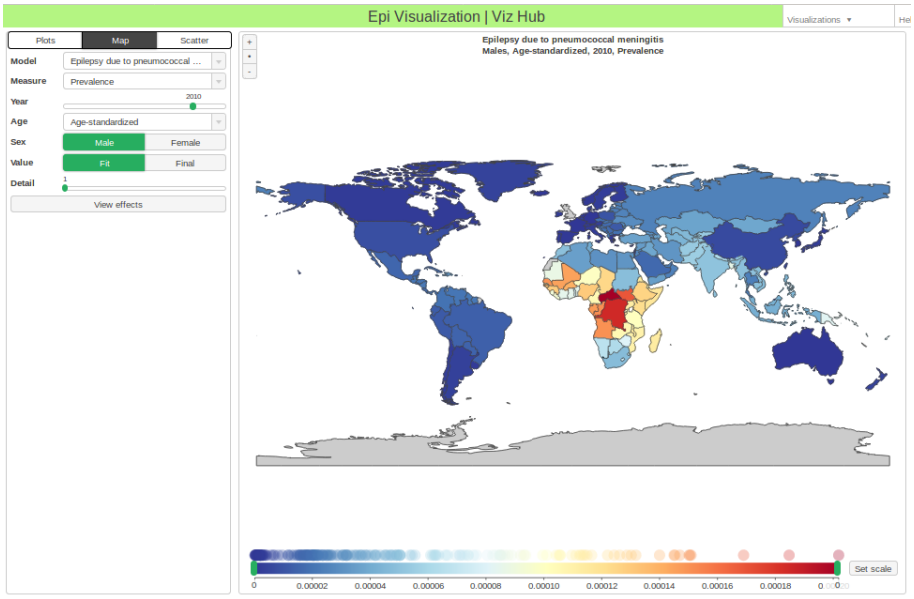


FIGURE 4.25 – Un exemple de visualisation de l'épilepsie due à la méningite H influenza de type B dans l'année 2010 avec Epi Visualization.

4.2.3 Discussion

La plupart des visualisations décrites ci-dessus respectent les critères de base; elles répondent donc aux questions : **où**, **quand** et **quoi**. Cependant, la manière dont ils sont traités varie d'une visualisation à une autre.

L'information associée à une situation géographique (C1) est représentée de différentes manières, principalement en fonction de l'origine des sources (données statistiques vs. données précises). Par exemple, Epi Visualization, via la carte choroplèthe, se concentre sur la visualisation de données dans des régions spécifiques. HealthMap

utilise une carte de chaleur pour tenir compte de la densité de données. Des approches offrent des positions exactes des données comme Empress-i et GapMinder. Il faut noter toutefois que les approches qui possèdent des informations précises peuvent aussi offrir des méthodes d'agrégation d'informations, notamment sous la forme de cartes de chaleur ou de cercles de rayons variables comme nous le précisons ultérieurement.

Le temps (C2) est traité de différentes manières (statiques ou dynamiques) : la liste (Empress-i, HealthMap), le curseur (Epi Visualization, GapMinder) ou le diagramme en ligne (HealthMap). En fonction des applications, l'information est représentée précisément (GapMinder, Epi Visualization) ou bien via des résumés s'il s'agit de données sources déjà agrégées (Empress-i, HealthMap).

Les différents systèmes utilisent généralement une ou plusieurs sources qui peuvent être de différents types (C3) : officielles, non officielles (généralement des textes déposés par des experts ou obtenus à l'aide d'approches de fouille de textes), statistiques, etc. Ces sources peuvent contenir diverses catégories d'informations (e.g. maladies, espèces, symptômes, etc.) et seules deux d'entre elles peuvent en représenter plusieurs. HealthMap montre les espèces à l'aide d'une carte de chaleur et les maladies avec des cercles sur la carte. Empres-i représente les types de données en utilisant deux glyphes différents (cercle, triangle) et des couleurs sur la carte.

Les critères transversaux sont également traités de façon différente selon les visualisations. Les résumés d'informations (C4) sont abordés de différentes manières :

1. le résumé de **où** (C1) : par des cartes de chaleur (HealthMap, Empres-i), par des couleurs de régions dans les cartes choroplèthes (GapMinder, IHME) et par des couleurs de cercles dans les cartes de points (GapMinder).
2. le résumé de **quand** (C2) : par des diagrammes en ligne (HealthMap) et par des diagrammes en barres (Empres-i).
3. le résumé de **quoi** (C3) : par des tailles de cercles (GapMinder), par des dégradations de couleurs (Epi Visualization, HealthMap) et par des utilisations de formes différentes (Empres-i).

Les interactions (C5) aussi sont spécifiques à chaque visualisation, et sont liées aux critères de base :

1. le **où** (C1) : choix à partir d'une liste et par continent (GapMinder), sélection par forme rectangulaire sur la carte (Empres-i), choix manuel (HealthMap), choix par région (IHME).
2. le **quand** (C2) : choix à partir d'une liste (Empres-i), choix manuel (HealthMap), choix par curseur (GapMinder, Epi Visualization).
3. le **quoi** (C3) : sélection de cercles (GapMinder, HealthMap, Empres-i), sélection des informations à partir d'une liste (Empres-i, HealthMap), sélection d'une valeur sur un ascenseur (Epi Visualization).

Enfin, la représentation des sources est présente (C6) mais n'est pas forcément explicite visuellement. GapMinder utilise une liste et HealthMap des petites icônes.

L'utilisateur ne peut donc pas directement voir l'origine sur la carte car elles possèdent souvent la même représentation. L'information précise n'est obtenue qu'à l'aide de mécanismes d'interaction.

La Table 4.1 résume les différences entre les visualisations présentées par le biais des critères préalablement définis.

Outils	Les critères de bases				Les critères transversaux						
	C1	C2	C3		C4			C5			C6
			Sources	Types	Où	Quand	Type	Où	Quand	Type	
Empres-i	Oui	Oui	Résumé de rapports	Maladies/espèces	Oui	Oui	Oui	Oui	Oui	Oui	Non
HealthMap	Oui	Oui	Articles/Données statistiques	Maladies/espèces	Oui	Oui	Oui	Oui	Oui	Oui	Partiel
GapMinder	Oui	Oui	Données statistiques	Maladies	Oui	Non	Oui	Oui	Oui	Non	Partiel
Epi Visualization	Oui	Oui	Données statistiques	Maladies	Oui	Non	Oui	Oui	Oui	Oui	Non

TABLE 4.1 – Comparaison des différentes approches de veille épidémiologique au regard des différents critères.

Il existe aussi d'autres approches qui ne traitent pas d'épidémies mais qui permettent néanmoins de visualiser des informations spatio-temporelles qui satisfont les critères définis précédemment. Par exemple, le système VaiRoma [20] proposé par Cho et al. est un système dédié à l'histoire romaine. Les données sont notamment issues de Wikipédia à l'aide d'algorithmes de fouille de textes. Les informations spatio-temporelles sont visualisées afin de permettre à un utilisateur de naviguer et de mieux appréhender les événements de l'histoire romaine. L'outil contient de nombreuses vues : temporelle, géographique, circulaire, et arbre de sujets. L'utilisateur peut sélectionner un intervalle de temps (C5) (Cf. Figure 4.26). Les données associées sont alors affichées sur la carte sous forme de carte de chaleur afin de mettre en évidence la répartition géographique des données. L'utilisateur peut également sélectionner des sujets spécifiques qui apparaîtront précisément sur la carte.

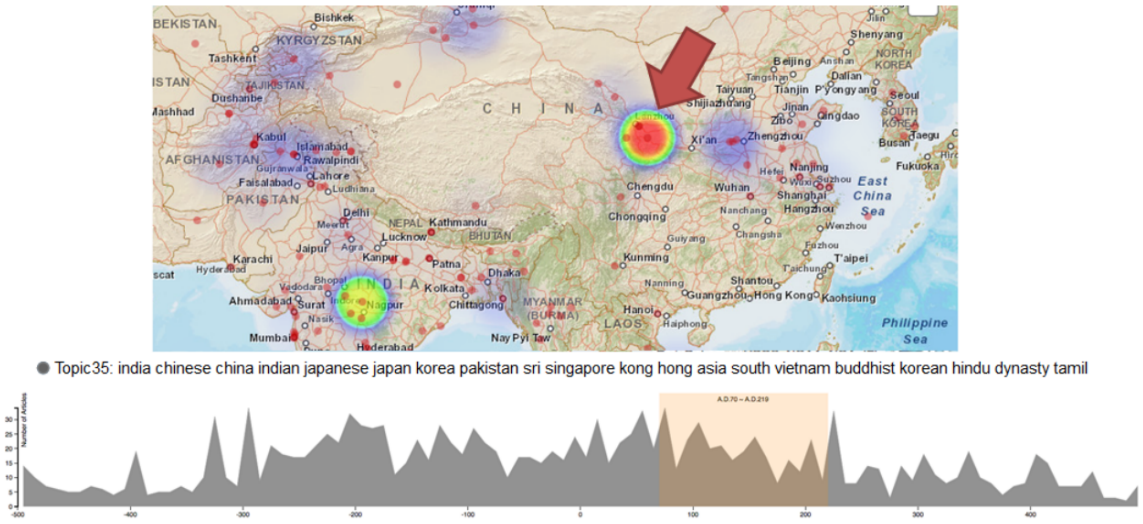


FIGURE 4.26 – La sélection de la période allant de l'an 70 à l'an 219 sur le sujet 36 conduit à des points colorés localisés en Chine pendant la dynastie des Han (flèche rouge).

La Figure 4.27 montre la vue circulaire des sujets : le Sunburst au centre représente la structure hiérarchique de ces sujets. Le texte au centre correspond soit à la plage de temps actuellement sélectionnée, soit aux mots-clés du sujet (voir rectangle rouge). Chaque cercle extérieur représente un sujet. Un zoom permet de lire plus facilement les mots-clés de ce sujet (rectangle orange).

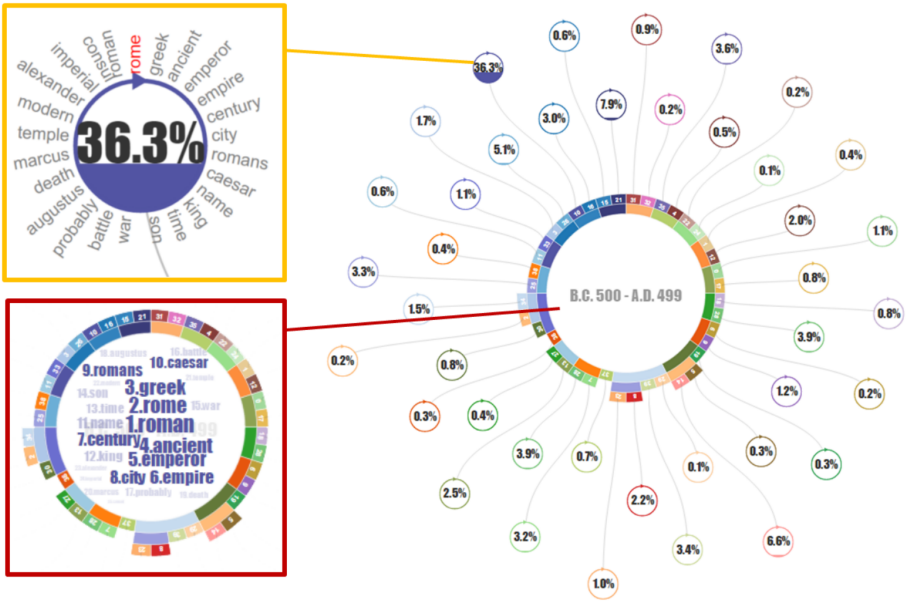


FIGURE 4.27 – Vue circulaire des sujets.

VaiRoma permet à l'utilisateur de comparer des données issues d'intervalles de temps différents. Par exemple, la Figure 4.28 montre une comparaison de données spatio-temporelles entre -220 et -201 (gauche) et entre -60 et -41 avant J.C. (droite) sur le sujet 16 (guerre dans Rome). Les 3 principaux articles de Wikipédia de la vue de gauche sont "Bataille de Causium", "Ra Valley Raid" et "Bataille de Lilybeaum". Les 3 principaux articles de la vue de droite sont "Bataille d'Alésia", "Bataille d'Utica (49 av. J.C.)" et "Guerres gauloises".

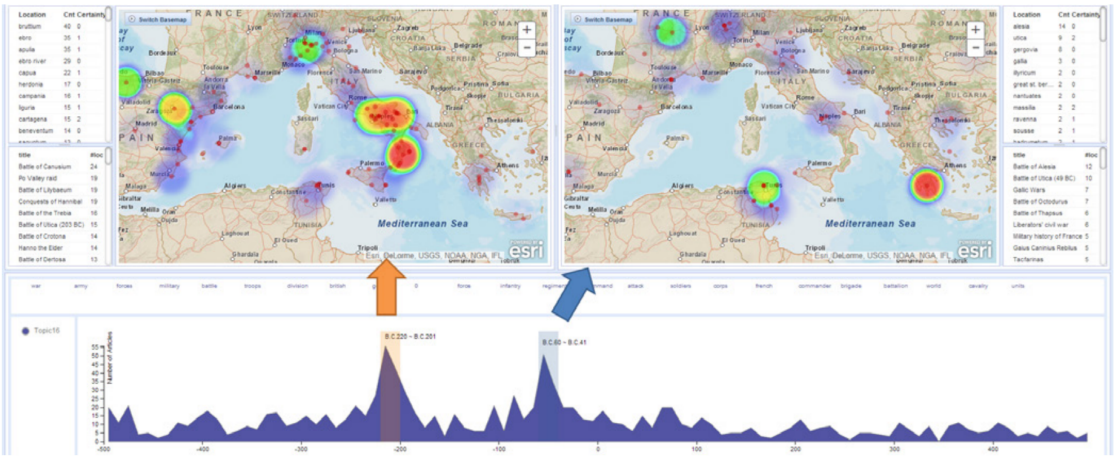


FIGURE 4.28 – Comparaison des périodes -220 à 201 av. J.C. (gauche) et -60 à -41 av. J.C. (droite) sur le sujet 16 (guerre).

En se basant sur les critères définis avec les experts et sur l'étude de l'état de l'art pour la visualisation d'information spatio-temporelle, nous proposons une nouvelle approche, EpidNews, que nous allons maintenant présenter.

4.3 EpidNews

EpidNews[11] est un nouvel outil de visualisation analytique pour le suivi de données sources liées à l'épidémiologie animale. Ce système prend en entrée deux types de données : (1) des données épidémiologiques (lieux, dates, symptômes, etc.) issues de sources officielles, (2) les mêmes types d'informations extraites par des approches de fouille de textes [4, 7]. Les documents (c'est-à-dire, dépêches) associés à ces informations sont également disponibles et seront prises en compte par EpidNews. Il satisfait tous les critères présentés préalablement (voir la section 4.2.1). Il est composé de plusieurs vues interactives (C5) qui permettent aux utilisateurs d'analyser, d'observer et de suivre facilement les informations spatio-temporelles des épidémies provenant de multiple sources (C6).

La Figure 4.29 présente une vue générale d'EpidNews. Cette vue est composée : 1) d'une carte géographique (Figure 4.29.a) qui montre les informations spatiales (C1), 2) de deux streamgraphs (Figure 4.29.b) pour exprimer les informations temporelles des données officielles, i.e. issues de sources officielles internationalement reconnues, et non officielles, i.e. issues d'articles de presse, (C2), et 3) d'une représentation de type sunburst (C4) (Figure 4.29.c) qui présente la relation entre les différents types de données (les maladies, les espèces et les symptômes) de manière hiérarchique. Chaque type de données contient un ensemble d'entités (C3) : par exemple, le type "maladies" contient *African swine fever, Avian Infuenza, Bluetongue, Foot-and-mouth-disease*, etc. Ces informations sont exprimées dans le gestionnaire de données (Figure 4.29.d) qui permet à l'utilisateur de sélectionner et de filtrer les données selon ses besoins.

4.3.1 Carte géographique

Nous visualisons l'information spatiale (C1) à l'aide d'une carte interactive (Cf. Figure 4.29.a) qui permet d'offrir à l'utilisateur la possibilité d'appréhender les données selon deux approches : 1) de manière résumée en montrant la densité des entités dans des zones géographiques à l'aide d'une carte de chaleur, et 2) de manière plus détaillée en plaçant les entités à leur position géographique précise à l'aide d'icônes.

L'utilisation de la carte de chaleur (Cf. Figure 4.30) permet à l'utilisateur de rapidement évaluer la densité d'articles dans une zone géographique donnée. La variation de densité est prise en compte en utilisant une échelle de couleurs qui va du bleu (faible densité) au rouge (forte densité). L'utilisateur peut ainsi facilement repérer les zones qui ont plus d'entités (ou plus denses) par rapport aux autres zones.

11. https://youtu.be/N8yfM42P4ME

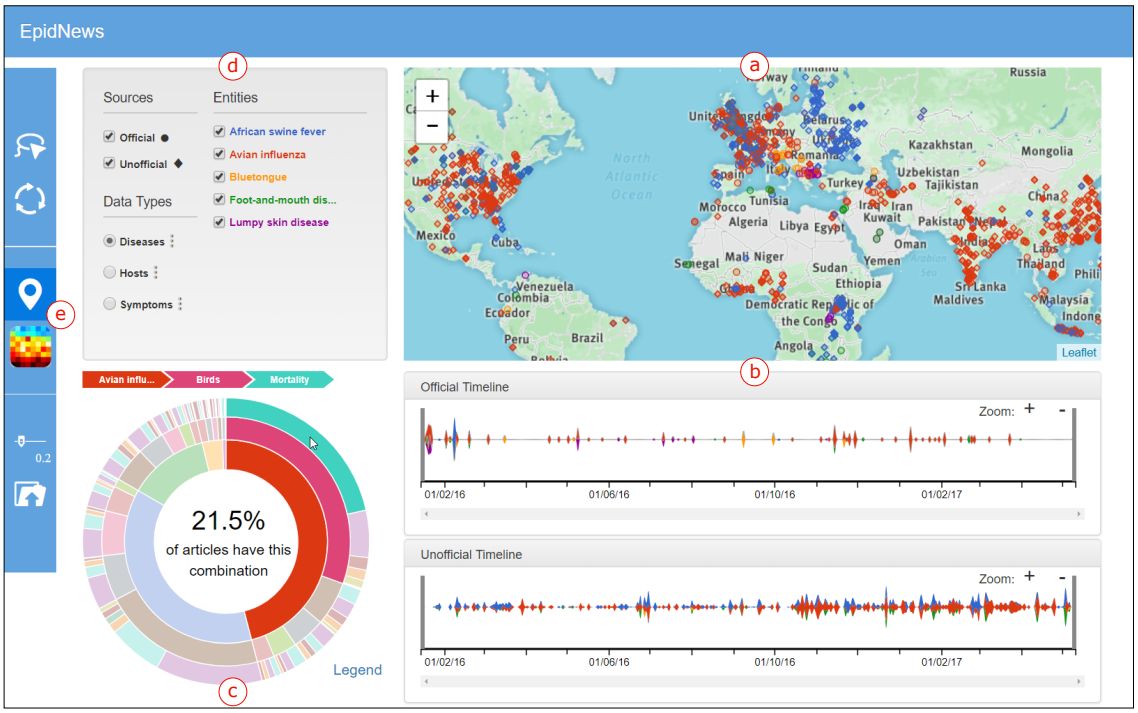


FIGURE 4.29 – Vue générale d'EPIDNEWS. **(a)** Une carte montre les emplacements des épidémies en utilisant un cercle ou un diamant selon le type de source. **(b)** Deux streamgraphs permettent de comparer l'évolution temporelle des sources officielles et non officielles. **(c)** Un sunburst présente les relations entre les maladies, les espèces et les symptômes dans une vue hiérarchique. **(d)** Un gestionnaire de données permet la manipulation des données représentées dans les autres vues (sources, type de données et entités). **(e)** Une barre d'outils offre d'autres fonctionnalités interactives.

Dans le cas où les articles sont directement affichés sur la carte, des glyphes différents sont utilisées pour représenter l'origine des sources : des cercles pour les sources officielles et des diamants pour les non officielles (Cf. Figure 4.29.a). Cette représentation facilite la distinction visuelle de chaque type de sources dans la carte (C6). En outre, des couleurs différentes sont attribuées aux glyphes pour décrire le type d'entité (Cf. Figure 4.29.d). Par exemple, un cercle rouge dans la Figure 4.29.a représentent la maladie "*Avian influenza*" issue d'une source officielle. Étant donné que de nombreux articles peuvent être positionnés au même endroit, les glyphes peuvent se chevaucher. Nous attribuons à chaque glyphe un degré d'opacité faible afin de permettre à l'utilisateur d'estimer le nombre d'éléments à un point donné.

Différentes interactions (C5) sont associées à cette carte comme les actions classiques de zoom avant/arrière ou de déplacement.

L'utilisateur peut cliquer sur un glyphe et une fenêtre apparaît. Elle contient la liste de tous les articles localisés sous le curseur de la souris (Cf. Figure 4.31). Chaque

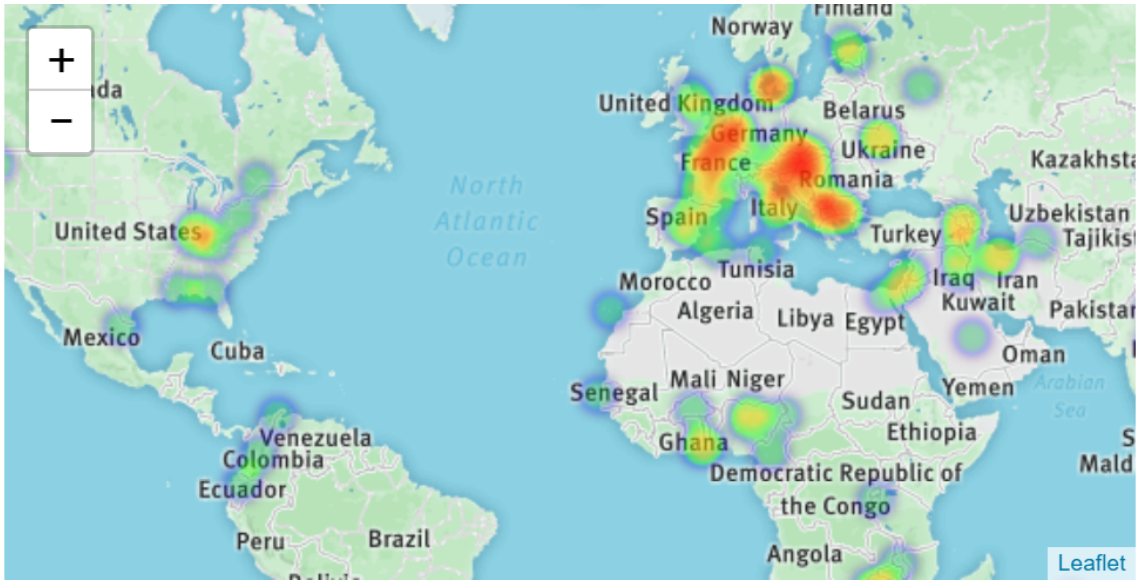


FIGURE 4.30 – Un exemple de carte de chaleur qui représente la distribution des maladies provenant de sources officielles.

élément de cette liste mentionne les informations suivantes : le type de source et l'entité (la forme et la couleur du glyphe), la date, un indice de confiance (une mesure sur la précision de la géolocalisation de l'article), et un lien hypertexte vers la source de l'article. L'utilisateur peut filtrer un type de source ou certaines entités particulières en décochant les cases correspondantes dans les sections "Sources" et "Entities" du gestionnaire de données (Cf. Figure 4.29.d). Par conséquent, les données de toutes les vues sont mises à jour simultanément.

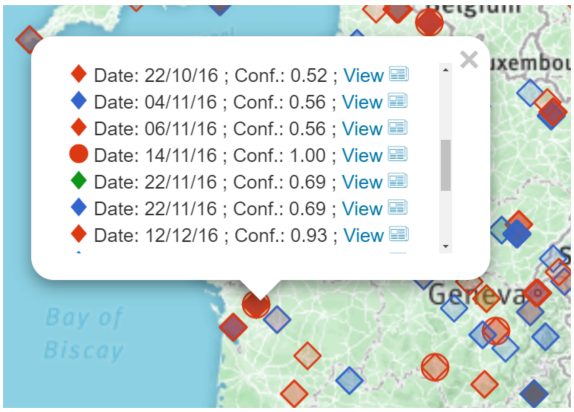


FIGURE 4.31 – Informations officielles (cercles) et non officielles (diamants) au sujet de maladies localisées près de Bordeaux, en France : (*avian influenza* en rouge, *African swine fever* en bleu et *foot and mouth disease* en vert).

Enfin, l'utilisateur peut basculer de la vue avec carte de chaleur à la vue avec glyphes, et vice versa, en cliquant sur le troisième ou le quatrième bouton de la barre d'outils (Figure 4.29.e).

4.3.2 Streamgraphs

Dans la mesure où nous disposons de deux types de données sources, nous représentons l'évolution temporelle (C2) du nombre d'articles (C3) à l'aide de deux streamgraphs [16] (Cf. Figure 4.29.b) : le premier contient des données provenant de sources officielles et le second de sources non officielles. Les deux streamgraphs ont le même intervalle de temps pour permettre à l'expert de comparer les informations des sources qui sont positionnées aux mêmes dates.

Les couleurs des entités (C4) sont les mêmes que dans le gestionnaire de données (Cf. Figure 4.29.d). À l'aide de ce dernier, l'utilisateur peut filtrer les articles dont le nombre est représenté par les streamgraphs selon leur type d'entité (espèces/maladies/symptômes) ou selon les entités qu'ils mentionnent, en décochant les cases correspondantes. Les deux streamgraphs sont alors mis à jour.

Les deux streamgraphs sont synchronisés (C5). Si l'utilisateur passe la souris sur l'un d'entre eux, les informations (le nombres d'articles à la date sélectionnée) apparaissent dans les deux streamgraphs. Par exemple, dans la Figure 4.32, le fait de survoler le stream rouge de données officielles (streamgraph du haut) a pour conséquence d'afficher les informations dans les deux vues. Elles montrent qu'il y avait 10 articles officiels sur l'*Avian influenza* le 14 novembre 2016, alors qu'il n'y en avait que 4 non officielles à la même date. Nous pouvons également observer que le stream survolé est mis en surbrillance dans les deux streamgraphs.

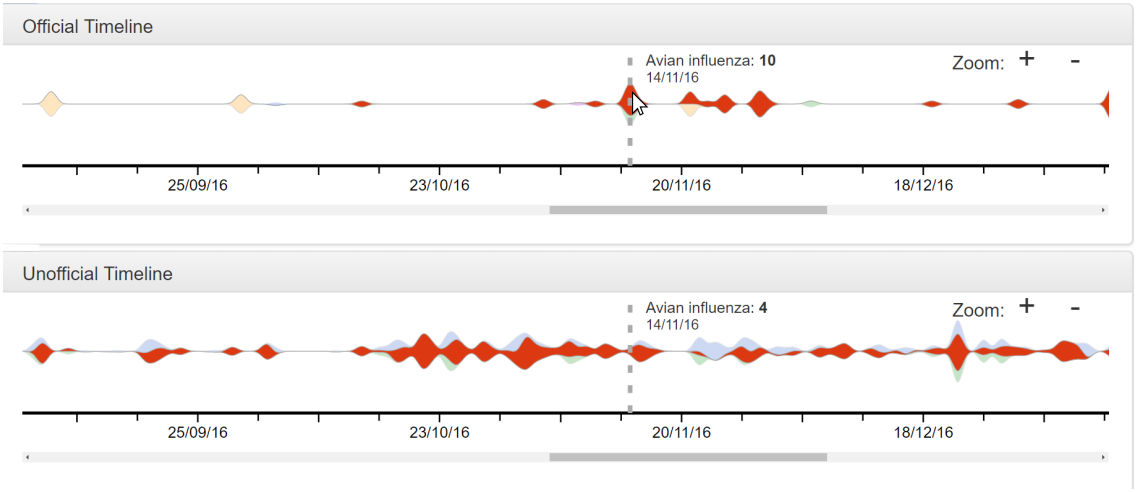


FIGURE 4.32 – Représentation et comparaison des informations temporelles à l'aide de streamgraphs : nombre d'articles parlant de l'*avian influenza* le 14 novembre 2016 (sources officielles en haut, sources non officielles en bas).

L'utilisateur peut modifier le niveau de zoom d'un streamgraph en cliquant sur les boutons dédiés placés en haut à droite de chacun d'eux. Un ascenseur horizontal en bas permet à l'utilisateur d'afficher les parties du streamgraph qui peuvent se retrouver en dehors de la vue lors d'un zoom avant. Les niveaux de zoom sont synchronisés dans les deux streamgraphs, si l'utilisateur augmente le zoom de l'un deux, il est aussi augmenté dans l'autre. Les deux ascenseurs horizontaux sont aussi synchronisés. De ce fait, la période couverte par les deux streamgraphs reste toujours la même, et nous évitons ainsi des erreurs de comparaison.

La barre d'outils (Figure 4.29.e) fournit une autre fonctionnalité interactive (C5) entre les streamgraphs et la carte. L'utilisateur peut activer un lasso en choisissant la première option dans cette barre, puis sélectionner une zone de forme personnalisée sur la carte. Les streamgraphs montrent alors la date des articles dans cette zone grâce à des lignes en pointillés verticales (Cf. Figure 4.32). Le deuxième bouton de la barre d'outils permet de réinitialiser le lasso.

Enfin, deux lignes verticales gris foncées sont initialement positionnées aux extrémités gauche et droite des streamgraphs (Cf. Figure 4.29.b). Elles peuvent être déplacées pour sélectionner les données dans une période de temps spécifique (Cf. Figure 4.33.b). La période sélectionnée (située entre ces lignes) est soulignée dans les streamgraphs, et les données des autres périodes sont supprimées des autres vues.

4.3.3 Sunburst

EpidNews permet de visualiser un résumé des informations spatio-temporelles (C4) à l'aide d'un sunburst [72]. Cette vue comprend trois niveaux hiérarchiques (trois anneaux) correspondant aux trois types d'entités (C3). Elle permet de bien montrer les relations entre ces types, à savoir les maladies, les espèces et les symptômes.

Par exemple, sur la Figure 4.33.c, l'anneau intérieur représente les espèces, l'anneau du milieu représente les symptômes et l'anneau extérieur représente les maladies. Les anneaux sont divisés en arcs, dont chacun représente une entité du type correspondant. Les couleurs des entités sont les mêmes que dans les autres vues. Via le sunburst, les utilisateurs peuvent observer le pourcentage de différentes combinaisons (C5) en passant la souris sur les arcs (entités). Par exemple, la Figure 4.33.c montre que 30% des articles évoquent de la combinaison *birds* (espèce), *mortality* (symptôme) et *avian influenza* (maladie).

L'utilisateur peut également cliquer sur une entité pour mettre en évidence dans les autres vues les données la concernant. Par exemple, dans la Figure 4.33, l'utilisateur a cliqué sur *avian influenza* : les informations correspondantes apparaissent sur la carte et les streamgraphs, les autres informations sont retirées de la carte et sont affadies dans les streamgraphs. Si l'utilisateur clique sur *mortality* (i.e. une entité de l'anneau du milieu), les informations combinant *birds* (anneau central) et *mortality* seront mises en évidence sans tenir compte de contraintes sur les entités de l'anneau extérieur (i.e. les maladies).

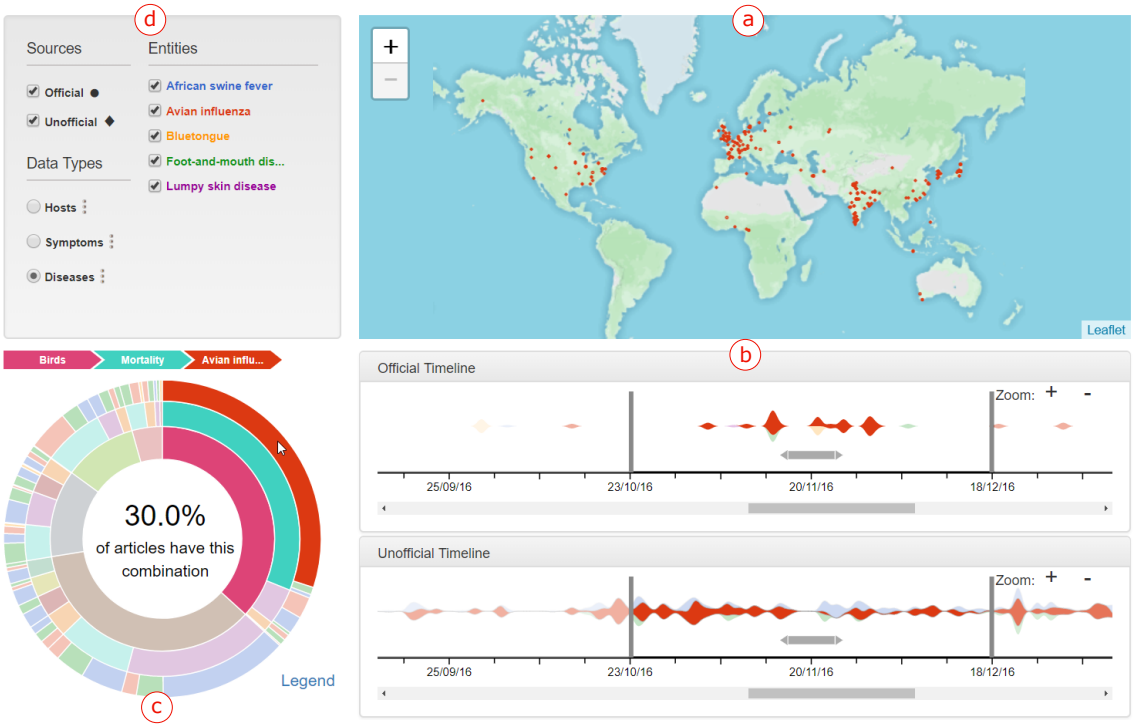


FIGURE 4.33 – Synchronisation entre les différentes vues d'EpidNews: **(a)** la carte, **(b)** les streamgraphs et **(c)** le sunburst.

Les utilisateurs ont également la possibilité de modifier l'ordre des niveaux dans la hiérarchie en faisant glisser les types d'entité dans le gestionnaire de données (C5). Par exemple, les types de données de la Figure 4.29.d et de la Figure 4.33.d ont un ordre différent. Les vues sont alors automatiquement mises à jour. Par exemple, le sunburst dans la Figure 4.29.c est ordonné de la façon suivante : les maladies apparaissent dans l'anneau interne, les espèces apparaissent dans l'anneau du milieu et les symptômes apparaissent dans l'anneau externe.

Comme nous l'avons décrit précédemment, la sélection d'une période dans les streamgraphs (en utilisant les deux lignes verticales grises) modifie les données visualisées dans le sunburst et la carte, en filtrant les éléments en dehors de cette période.

Enfin, concernant les deux derniers boutons disponibles dans la barre d'outils (Cf. Figure 4.29.e), le cinquième est utilisé pour définir le niveau de confiance minimal [12] pour l'emplacement des articles. De ce fait, les articles ayant une précision de localisation inférieure à la valeur sélectionnée ne sont pas visualisés sur la carte. Par exemple, dans la Figure 4.29, tous les articles avec une confiance d'au moins 0,2. Le dernier bouton permet quant à lui aux utilisateurs de charger de nouvelles données.

12. Confiance associée à la prédiction des entités extraites par les approches de fouille de textes [4].

4.4 Étude de cas

L'objectif de cette section est de voir si les fonctionnalités proposées dans EpidNews sont adaptées à la surveillance de la propagation des épidémies. Dans ce cadre, une étude de cas a été réalisée par une experte [13] en épidémiologie qui a choisi de mener une analyse portant sur l'*African swine fever* (*ASF*) et l'espèce *porcine* (*wild boars* et *domestic pigs*).

Pour cette étude, les données officielles proviennent des bases de données FAO [14] ("Organisation des Nations Unies pour l'alimentation et l'agriculture") et ADNS [15]. Les données non officielles sont obtenues en utilisant un outil de biosurveillance PADI-web [4] qui suit en ligne les articles de presse sur les épidémies. L'étude porte sur la période 10/02/2016 - 10/11/2017.

Trois tâches principales ont été définies :

— **Tâche 1 :** l'objectif est d'évaluer la facilité d'utilisation d'EpidNews pour la visualisation et l'analyse des informations officielles,

— **Tâche 2 :** l'objectif est d'évaluer l'intérêt d'EpidNews pour l'analyse de données issues de sources non officielles, à savoir la visualisation des indicateurs épidémiologiques (maladies, hôtes, symptômes) contenus dans les articles,

— **Tâche 3 :** l'objectif est d'évaluer la capacité d'EpidNews à lier les méta-données (titre de l'article, lien vers l'article source, etc.) aux informations géographiques.

4.4.1 Tâche 1 : Visualiser et analyser les données officielles

Pour la première tâche, l'experte a sélectionné la maladie *ASF*. Ensuite, elle a sélectionné une période de temps sur les streamgraphs et a déplacé cette période afin d'observer les données à différents moments. Elle a utilisé la carte de chaleur (Cf. Figure 4.34) pour distinguer les zones à forte densité (en rouge) et les zones à faible densité (en bleu). Cela a permis à l'experte de comprendre l'évolution de l'*ASF* en Europe et plus particulièrement de mieux comprendre comment cette épidémie s'est propagée de l'Europe de l'Est à l'Europe Occidentale. Elle a pu observer qu'il y a de moins en moins d'articles sur le sujet en Ukraine alors qu'il y en a de plus en plus en République tchèque.

Pour comparer les données des espèces touchées par l'*ASF*, l'experte a ensuite utilisé la vue détaillée avec les glyphes où les "*wild boars*" apparaissent en marron et les "*domestic pigs*" en rose (Figure 4.35). Ces couleurs rendaient les deux espèces facilement reconnaissables sur la carte. Elle a pu constater que les cas touchant les "*wild boar*" étaient principalement concentrés dans les pays baltes alors que ceux touchant les "*domestic pigs*" étaient répartis sur une zone plus vaste et diversifiée.

13. Sarah Valentin, CIRAD, ASTRE, TETIS, Montpellier
14. http://empres-i.fao.org/eipws3g/ (accédé le 14 mars 2018)
15. https://ec.europa.eu/food/animals/animal-diseases/not-system_en, (accédé le 14 mars 2018)

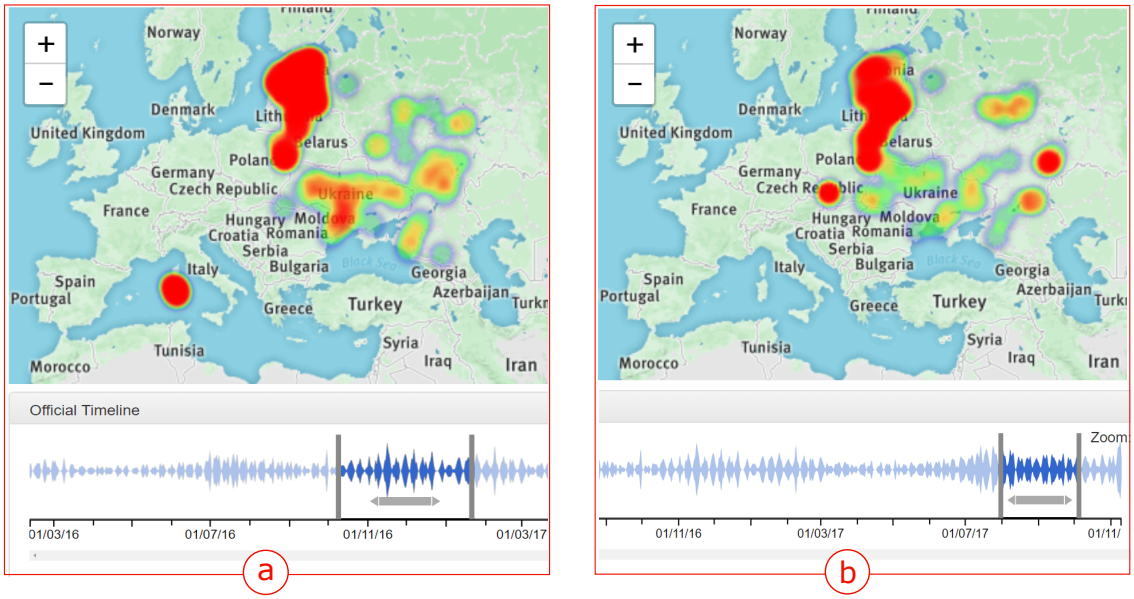


FIGURE 4.34 – Répartition des données officielles de l'*ASF* avec la carte de chaleur, (a) du 10/10/2016 au 20/01/2017 et (b) du 01/08/2017 au 05/10/2017.

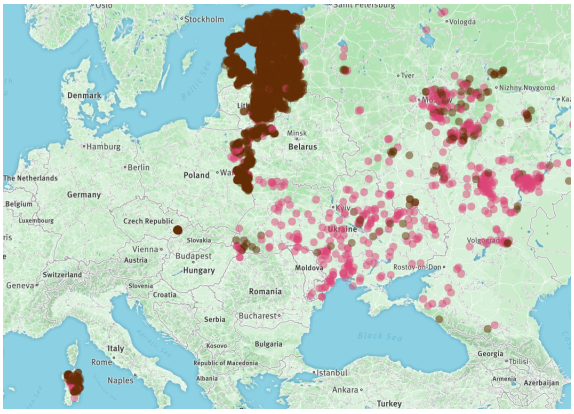


FIGURE 4.35 – Comparaison des espèces touchées par l'*ASF* dans les données officielles pendant toute la période étudiée : les articles mentionnant des "*wild boars*" sont représentés en marron et ceux mentionnant des "*domestic pigs*" en rose.

En observant les données officielles dans les Figures 4.34 et 4.35, une anomalie géographique potentielle a également été constatée : bien que constamment entourée de cas d' *ASF*, la Biélorussie n'a officiellement notifié aucun cas.

Le lasso a aidé l'experte à se concentrer sur des lieux particuliers indépendamment des frontières des pays. Les dates correspondant aux emplacements des données sont mises en évidence par des lignes pointillées dans le streamgraph. La sélection

d'un groupe de données dans l'ouest de la Russie, et l'observation des dates dans le streamgraph (Figure 4.36) suggèrent que ces épidémies d'*ASF* étaient non seulement liées dans l'espace mais aussi dans le temps : identification d'un lien épidémiologique probable entre les deux.

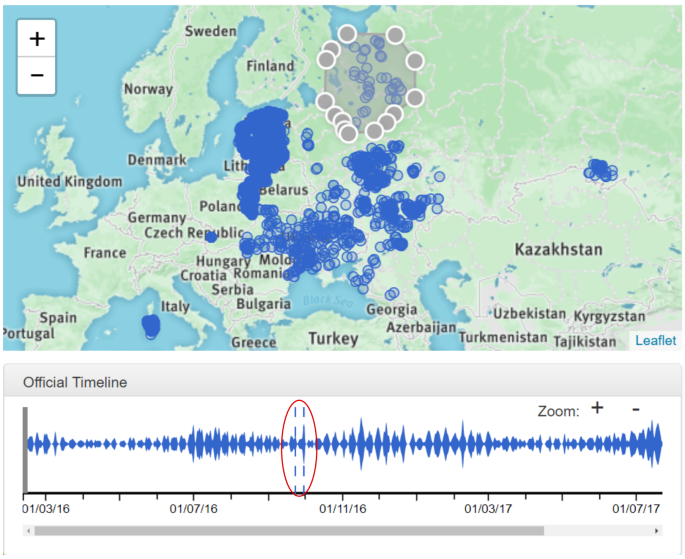


FIGURE 4.36 – Le streamgaph montre les dates des articles sélectionnés dans la carte à l'aide du lasso.

Pour certains pays, les articles parlant d'épidémies et ceux confirmant ces épidémies sont disponibles. Par exemple, pour la Russie (Figure 4.37) le stream d'*épidémie possible* en rouge et le stream de *confirmation d'épidémie* en bleu suggère qu'un groupe de cas survenus à la mi-septembre 2016 a été signalé avant d'être confirmé trois semaines plus tard.

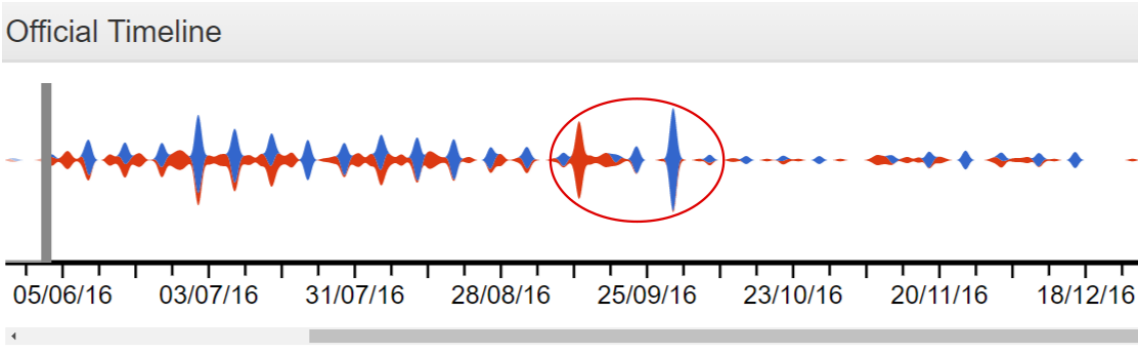


FIGURE 4.37 – Le stream d'épidémies possibles (en rouge) et le stream de notifications/-rapports (en bleu) concernant l'*ASF* en Russie.

4.4.2 Tâche 2 : Analyser différents types de données

Certaines informations non-officielles peuvent contenir des données qui ne sont pas spécifiques à la maladie concernée (c-à-d, des informations sur d'autres maladies/espèces/symptômes). Par exemple, 76,5% des documents non officiels comportent des informations combinant à la fois *ASF* et *domestic pigs* ou *wild boars*. Ceci est normal dans la mesure où il s'agit d'une maladie spécifique à cette espèce. Aussi, les 23,5% restants concernent d'autres combinaisons de maladies et d'espèces.

Après avoir évalué la proportion de chaque combinaison de maladies, d'espèces et de symptômes (Figure 4.38.b), l'experte a pu constater que *fever* est le symptôme le plus commun associé à l'*ASF*.

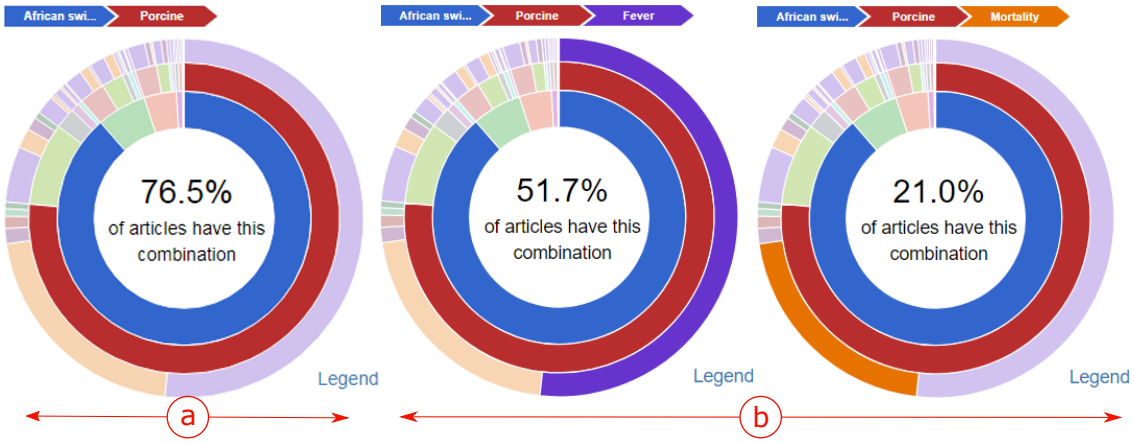


FIGURE 4.38 – **(a)** Exemple de combinaison maladie-espèce dans les sources d'informations non officielles. **(b)** Exemple de combinaisons maladie-espèce-symptôme dans le même type de sources d'informations.

4.4.3 Tâche 3 : Aperçu des articles originaux

Après avoir observé la distribution géographique des données non officielles autour de la Biélorussie, l'experte a souhaité consulter en détail le contenu des articles. En cliquant sur les icônes et en utilisant la liste, elle a confirmé que le contenu des articles était pertinent pour l'*ASF* et la Biélorussie.

En fait, deux articles concernaient l'interdiction du porc imposée par la Biélorussie en raison d'épidémies d'*ASF* dans les pays limitrophes et de la création d'un centre d'élevage de porcs. Un autre rapport était un article d'actualité pertinent faisant référence à des épidémies non déclarées par des sources officielles (Figure 4.39).

4.4.4 Conclusion de l'étude de cas

EPIDNEWS s'est révélé être d'une grande aide pour analyser facilement les données. L'experte a trouvé très utile de pouvoir localiser des groupes de cas potentiels

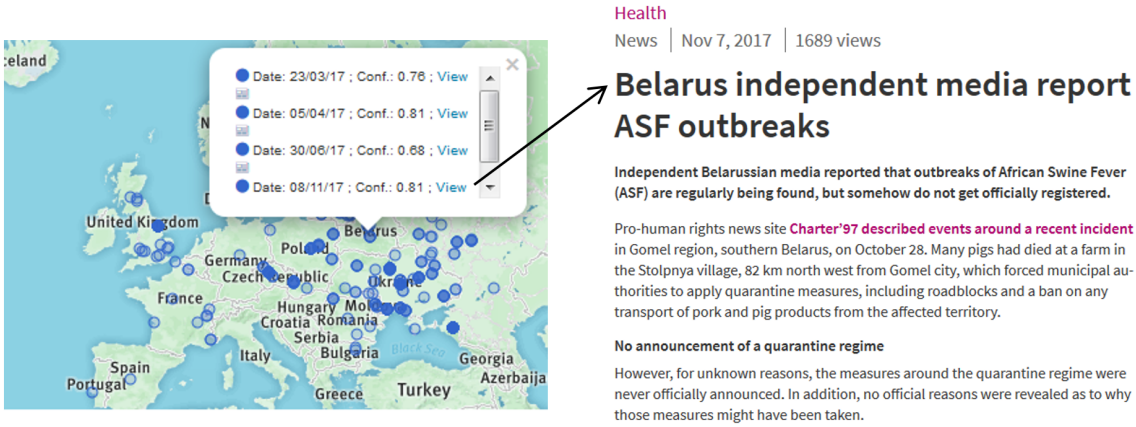


FIGURE 4.39 – Article sur l'*ASF* en Biélorussie.

avec un lien épidémiologique qui peut être validé avec des méthodes statistiques. L'outil a également aidé à détecter des biais dans l'ensemble de données, comme l'absence potentielle de notifications officielles dans certaines régions. Ainsi, EPIDNEWS et ses nouvelles fonctionnalités permettent de mettre en avant de nouvelles informations plus difficiles à obtenir avec les systèmes traditionnels, à savoir HealthMap [16] et Empres-i [24] (voir la section 4.2.2).

L'experte a précisé qu'EPIDNEWS était un outil efficace pour analyser des articles épidémiologiques et a particulièrement apprécié la combinaison de trois indicateurs importants en épidémiologie animale : les maladies, les espèces, et les symptômes.

Il est également utile pour détecter les ambiguïtés géographiques et les épidémies potentielles qui ne sont pas souvent détectées par les sources officielles. Elle a conclu que cet outil devrait être utilisé dans le suivi quotidien et l'analyse des sources d'information officielles et non officielles et qu'il réduit considérablement le travail manuel de l'équipe de surveillance épidémiologique.

4.5 Conclusion

Dans ce chapitre, nous avons présenté EPIDNEWS, un nouvel outil de visualisation analytique pour permettre d'aider les experts dans la surveillance d'épidémies animales. Il combine plusieurs vues complémentaires (carte géographique, streamgraphs et sunburst) pour explorer des données aussi bien officielles que non officielles. Des glyphes spécifiques aux différents types de données permettent de les localiser rapidement. Enfin, les interactions offrent de réelles facilités pour filtrer les données et mettre à jour automatiquement les vues. L'étude de cas réalisée par une experte en épidémiologie a montré que les fonctionnalités proposées par l'approche étaient effectivement adaptées à la surveillance de la propagation des épidémies.

16. https://healthmap.org

Nous revenons, à présent, sur certains choix liés notamment au chevauchement d'informations sur une carte. Considérons la carte de la Figure 4.40 dans laquelle les maladies sont représentées à l'aide de glyphes. Actuellement, nous considérons qu'une information sur un type de maladie apparaît sous la forme d'un cercle de couleur avec une certaine opacité. Pour afficher plusieurs informations localisées au même endroit, les glyphes sont superposées et l'opacité permet de donner une indication approximative sur leur nombre. Cette approche aide à résoudre le problème de la superposition d'information mais soulève un nouveau problème : comment savoir s'il y a 10 ou 100 objets superposés. Comme nous l'avons vu dans l'état de l'art, il existe différentes solutions pour représenter la densité. Les cartes choroplèthes ou les cartogrammes ne sont pas adaptés et nous avons choisi d'ajouter des cartes de chaleur pour visualiser les différentes densités (Cf. Figure 4.41). Cette solution, qui a d'ailleurs été appréciée par les experts qui peuvent rapidement repérer les zones d'intérêt et avoir une première approximation du nombre d'informations, a permis de résoudre partiellement le problème des chevauchements. Par contre, étant donnée l'hétérogénéité des données qui peuvent être localisées à un même endroit, il devient indispensable de poursuivre nos travaux afin de trouver une solution permettant de limiter ces chevauchements. Nous reviendrons sur ce point dans les perspectives de ce mémoire.

FIGURE 4.40 – Problème de surcharge de données sur la carte.

FIGURE 4.41 – La distribution des maladies officielles en utilisant la carte de chaleur.

Chapitre

5

Conclusions et perspectives

Sommaire

5.1 Synthèse des principales contributions

Dans cette section nous revenons sur les principales contributions présentées tout au long de ce mémoire.

Dans un premier temps, nous avons proposé un système de construction visuelle de requêtes pour la veille en épidémiologie animale EpidVis [36]. Il est composé de plusieurs vues : (1) un gestionnaire de mots-clés permet d'aider les épidémiologistes à représenter et structurer leurs connaissances (mots-clés et relations), (2) un constructeur de requêtes permet de construire et lancer automatiquement des requêtes, (3) une vue des résultats permet d'afficher les pages retournées par les moteurs de recherche et propose des fonctionnalités de filtrage et de sauvegarde, et (4) une visualisation de suggestions issues de fichiers externes permet d'enrichir les données du gestionnaire de mots-clés. Différentes évaluations ont été proposées : une étude utilisateur, une étude de cas et une discussion sur les requêtes complexes. L'étude utilisateur a montré que les différentes visualisations sont utiles et faciles à utiliser. L'étude de cas a permis de mettre en évidence qu'EpidVis offre une aide réelle aux experts dans leur tâche de veille. Enfin, la discussion sur les requêtes complexes a donné aux utilisateurs certaines informations sur le comportement des moteurs de recherche dans le cas de requêtes complexes.

La seconde contribution concerne la suppression de chevauchements [35]. Nous avons mis en évidence dans EpidVis que des chevauchements peuvent apparaître sur une dimension lors de l'opération de regroupement, i.e. lorsque les nœuds issus de la fusion sont positionnés au barycentre de l'ensemble des nœuds fusionnés. En généralisant le problème, nous avons mis en évidence 4 critères : la suppression de chevauchement; l'utilisation optimale de la longueur du segment; la préservation de l'ordre initial et la préservation des distances relatives. Certains de ces critères ont été abordés pour deux dimensions et le dernier critère est spécifique à une dimension et induit des propriétés visuelles intéressantes. L'algorithme proposé a une complexité temporelle de $O(|V|log(|V|))$. Deux études de cas basées sur des diagrammes en arcs, *les misérables* et un réseau de co-auteurs de PacificVis, ont été réalisées et ont mis en évidence l'importance du dernier critère pour faire apparaître des structures communautaires.

Dans la dernière contribution, nous avons proposé, EpidNews [44], un nouvel outil de visualisation analytique. Il permet d'explorer l'information spatio-temporelle à partir de différentes sources (officielles et non officielles) liées à l'épidémiologie animale. Différentes vues interactives ont été combinées pour gérer les besoins définis en collaboration avec des épidémiologistes. La carte permet de montrer les emplacements des épidémies en utilisant des glyphes adaptées en fonction des sources. Deux streamgraphs sont utilisés pour pouvoir comparer l'évolution temporelle des sources officielles et non officielles. Un sunburst met en évidence, pour un ensemble de sources sélectionnées, les relations entre les maladies, les espèces impactées et les symptômes observés. Un gestionnaire de données permet de manipuler les données représentées dans les autres vues (sources, type de données et entités). Enfin, des in-

teractions entre les différentes vues offrent des fonctionnalités pour filtrer les données représentées et mettre à jour l'ensemble des vues. Une étude de cas a été réalisée par une experte pour vérifier que les fonctionnalités proposées par EpidNews étaient adaptées à la surveillance de la propagation des épidémies. L'étude a montré qu'elle aidait à analyser le contenu des sources non officielles concernant la combinaison de trois indicateurs pertinents en épidémiologie animale (maladies/espèces/symptômes). Elle a aussi mis en évidence que la présentation des articles en fonction de leur contenu géographique était un moyen efficace de détecter les ambiguïtés géographiques et les foyers potentiels non détectés par les sources officielles. Enfin l'experte a conclu qu'un tel outil facilite grandement la surveillance et l'analyse quotidienne des sources officielles et non officielles.

5.2 Perspectives

5.2.1 Enrichir les connaissances dans le gestionnaire de mots-clés

La visualisation du gestionnaire de mots-clés actuelle est basée sur trois axes représentant les maladies, les symptômes et les hôtes. Ce choix de représentation a été fait suite à de nombreuses réunions avec les experts. L'objectif était d'offrir une visualisation adaptée aux besoins des utilisateurs en se focalisant sur les tâches qu'ils ont à accomplir. L'une des perspectives est maintenant de pouvoir étendre ces représentations.

La première possibilité serait de pouvoir ajouter des contraintes sur l'un des axes. Par exemple cela pourrait permettre de mettre en évidence des relations du type "espèces ayant une maladie" - "espèces transférant une maladie" (e.g. les relations entre *moustique* et *vache*). Ce type de contraintes peut être ajouté dans la visualisation des mots-clés et des suggestions mais soulève tout de même la question suivante : comment prendre en compte cette information sans trop surcharger la visualisation ?

Une autre possibilité serait également de pouvoir intégrer une ou plusieurs autres dimensions. Considérons, par exemple, la prise en compte d'une dimension localisation. Celle-ci pourrait avoir plusieurs usages : représenter dans le gestionnaire de mots-clés les lieux où les hôtes ont été infectés, faciliter la recherche de documents sur un lieu donné, etc. Bien entendu, en fonction de la problématique visée, les visualisations peuvent être très différentes. Restreindre une recherche de documents dans une région peut facilement se faire dans la visualisation de requêtes en intégrant une carte et en précisant, dans les critères, de prendre en compte la localisation. La visualisation de la localisation ou d'une autre dimension dans le gestionnaire de données engendrent de nouveaux problèmes : comment ne pas surcharger la représentation ? la représentation actuelle est-elle encore adaptée ? Comme nous l'avons vu dans l'état de l'art, il existe de nombreuses approches pour représenter les connaissances et, dans ce cadre, il est important de mieux définir les objectifs pour rechercher les

approches visuelles pertinentes. L'expérience acquise avec EpidVis nous a montré que la visualisation des mots-clés avait aussi une conséquence sur les requêtes qui peuvent être générées mais également sur le mécanisme de suggestion. Une nouvelle étude bénéficiera bien entendu de cette expérience.

5.2.2 Chevauchement de textes autour des cercles

Dans le cadre des suggestions, différents choix ont été réalisés pour retenir une visualisation adaptée aux besoins des utilisateurs. Il s'est avéré lors de l'utilisation d'EpidVis que des problèmes de chevauchements pouvaient apparaître (C.f. Figure 5.1).

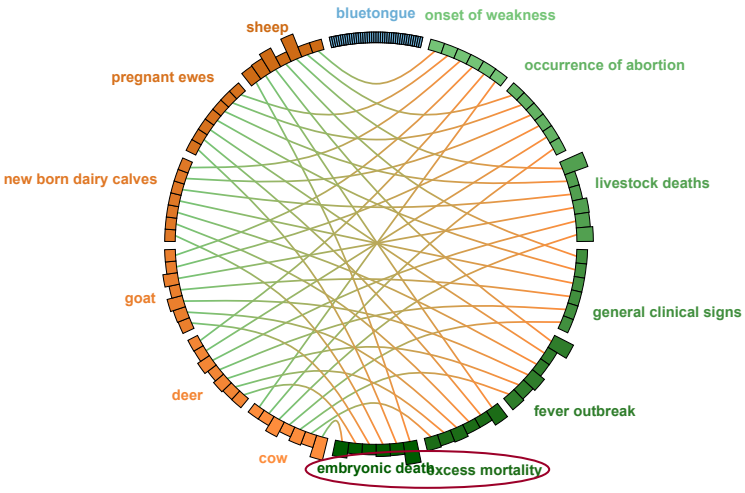


Figure 5.1 – Le chevauchement de textes dans la visualisation de suggestions.

Cette perspective concerne donc la suppression de chevauchement des textes autour des cercles. Dans ce cadre il convient tout d'abord de revenir sur les différentes manières de visualiser l'information autour d'un cercle. La Figure 5.2 présente les différentes approches de la littérature proposées. Les premiers travaux consisteront tout d'abord à mener une réelle évaluation utilisateurs de manière à déterminer l'approche la plus adaptée.

Dans un second temps il conviendra, à partir de la visualisation retenue, de proposer de nouvelles approches de suppression des chevauchements. Une première piste de recherche est l'utilisation des coordonnées polaires pour mieux positionner les textes. De manière à évaluer cette approche nous avons développé une plateforme d'expérimentation qui permet de mieux appréhender le problème et faciliter l'évaluation des algorithmes développés (C.f. Figure 5.3).

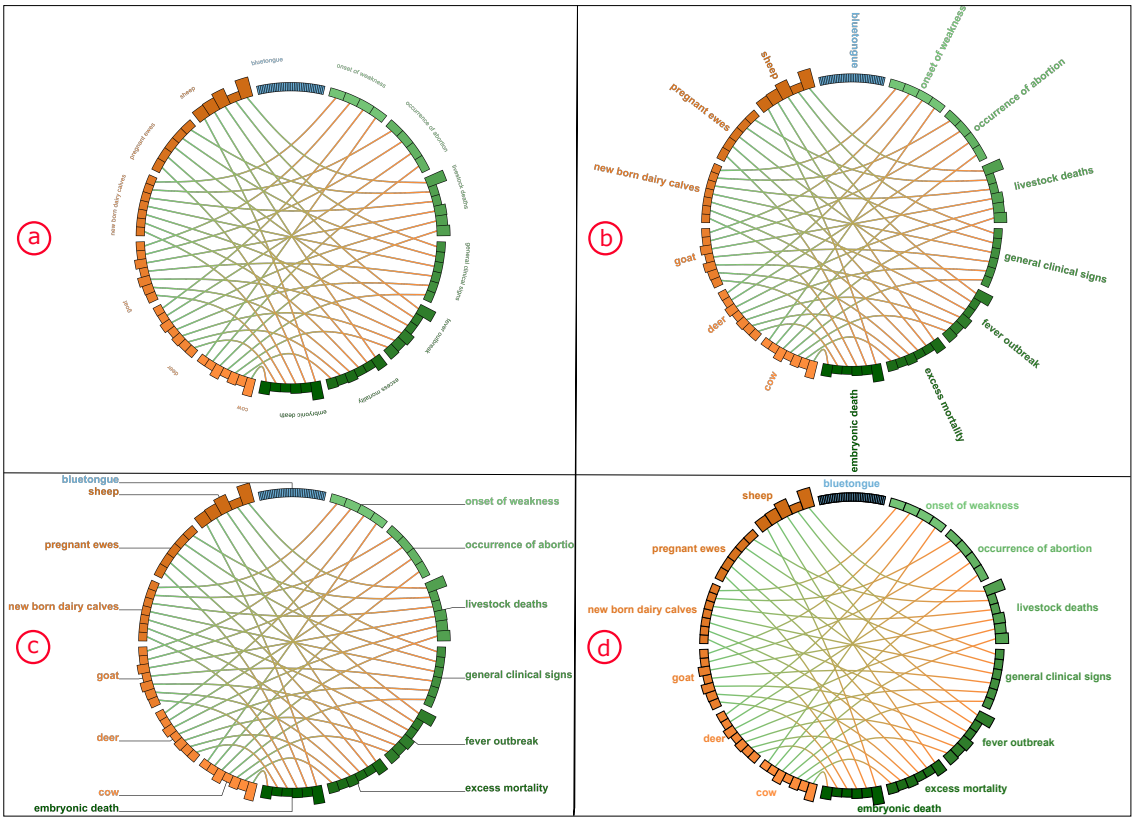


FIGURE 5.2 – Les différents types de visualisation de textes autour d'un cercle.

5.2.3 Visualisation de données hétérogènes sur une carte

La dernière perspective concerne la visualisation d'informations sur une carte. Comme nous pouvons le constater dans EPIDNEWS, de nombreuses informations hétérogènes sont placées sur la carte et nous avons développé différentes approches (glyphes, cartes de chaleur) pour pouvoir les mettre en évidence. L'ajout de nouvelles sources de données ou de nouvelles informations soulève un nouveau problème : comment représenter de nouvelles données dans la carte tout en permettant à l'utilisateur de les appréhender ? Cette problématique correspond, en fait, à un chevauchement très contraint : faire apparaître de très nombreuses informations dans une zone géographique. Actuellement deux grandes approches existent. Les premières suppriment des éléments. C'est par exemple le mécanisme utilisé dans Google Map où, en fonction du niveau de zoom, certains éléments disparaissent. Les secondes proposent d'agréger des éléments [81] comme l'illustre la Figure 5.4. Jusqu'à présent, les travaux se sont souvent focalisés sur le design des objets agrégés (par exemple utilisation de glyphes différents) et utilisent des algorithmes d'agrégation ou de clustering assez naïfs. Il conviendrait de mener de nouvelles recherches pour répondre aux questions suivantes : quels sont les critères d'agrégation utiles aux experts ?,

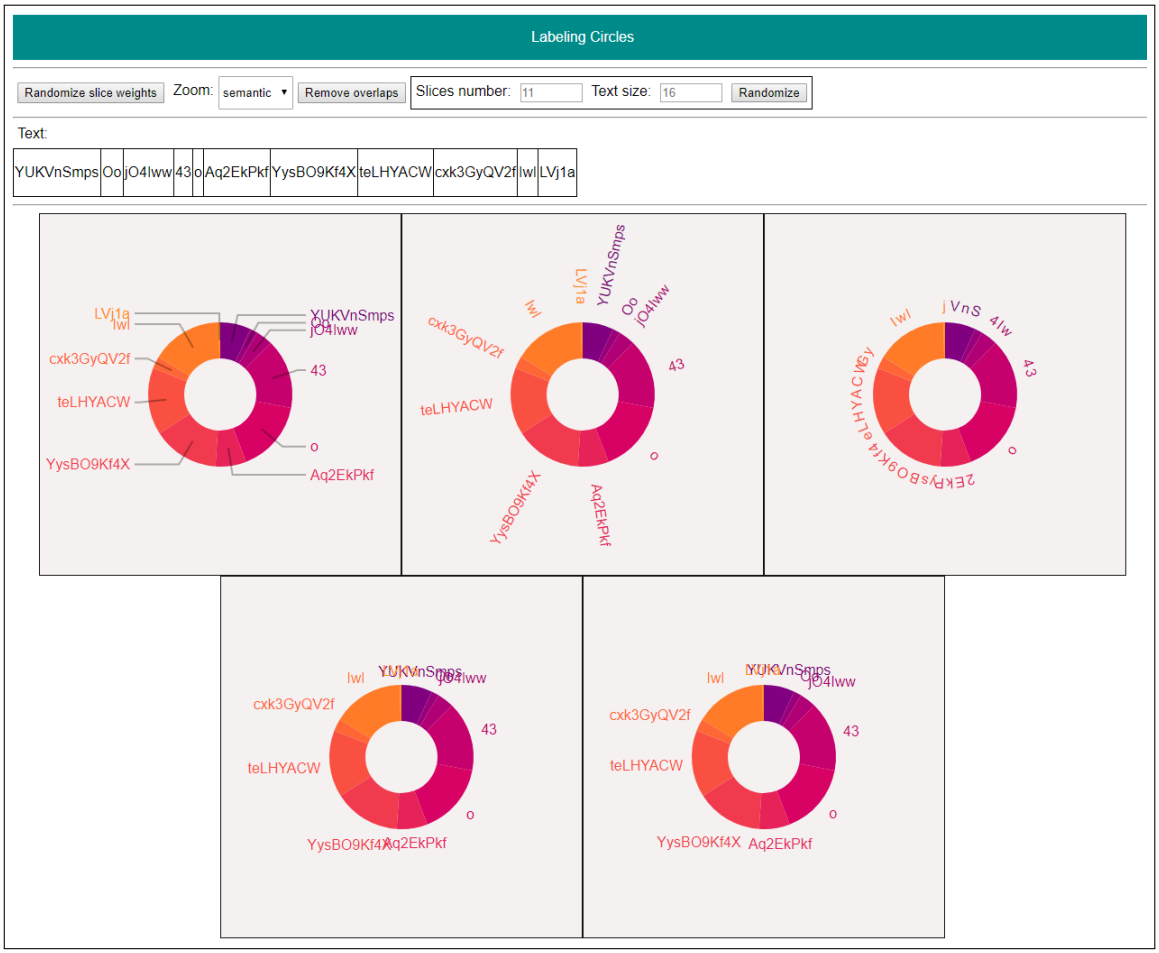


FIGURE 5.3 – Plateforme d'évaluation de suppressions de chevauchements.

comment minimiser le nombre d'agrégats ?, comment uniformiser la taille des agrégats ?, existe-t-il des types agrégations plus adaptées à des problématiques données et dans ce cas, est-il possible de proposer les différentes approches possibles ?

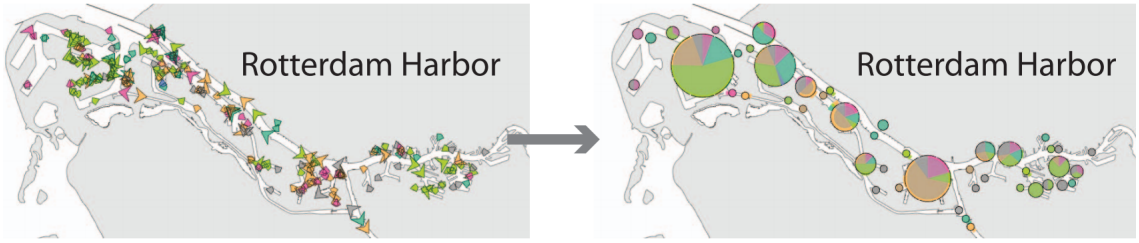


FIGURE 5.4 – Une approche d'agrégation de données.

Bibliographie

[1] Noboru Abe, Hiroaki Oh, and Kouhei Inoue. Algorithms for removing node overlaps with some basis nodes. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pages 93–102. Springer, 2016. (Cité aux pages 47, 63, 64 et 65.)

[2] Rakesh Agrawal, Behzad Golshan, and Evangelos E. Papalexakis. Overlap in the web search results of google and bing. *J. Web Science*, 2(2):17–30, 2016. (Cité à la page 38.)

[3] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of time-oriented data*. Springer Science & Business Media, 2011. (Cité à la page 71.)

[4] Elena Arsevska, Sylvan Falala, Jocelyn De Goer, Renaud Lancelot, Julien Rabatel, and Mathieu Roche. Padi-web: Platform for automated extraction of animal disease information from the web. In *Proceedings of Language and Technology Conference*, page 241–245, 2017. (Cité aux pages 92, 97 et 98.)

[5] Elena Arsevska, Mathieu Roche, Pascal Hendrikx, David Chavernac, Sylvain Falala, Renaud Lancelot, and Barbara Dufour. Identification of associations between clinical signs and hosts to monitor the web for detection of animal disease outbreaks. *International Journal of Agricultural and Environmental Information Systems*, 7(3):1–20, 2016. (Cité aux pages 14, 18 et 26.)

[6] Elena Arsevska, Mathieu Roche, Pascal Hendrikx, David Chavernac, Sylvain Falala, Renaud Lancelot, and Barbara Dufour. Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture*, 123:104–115, 2016. (Cité à la page 68.)

[7] Elena Arsevska, Sarah Valentin, Julien Rabatel, Jocelyn de Goer, Sylvain Falala, Renaud Lancelot, and Mathieu Roche. Padi-web dataset manually evaluated [1st january - 28th june 2016] cirad dataset, 2018. (Cité à la page 92.)

[8] Hiteshwar Kumar Azad and Akshay Deepak. Query expansion techniques for information retrieval: a survey. *Cornell University Library*, 2017. (Cité à la page 14.)

[9] Ziv Bar-Yossef, Idit Keidar, and Uri Schonfeld. Do not crawl in the dust: different urls with similar text. *ACM Transactions on the Web*, 3(1):3, 2009. (Cité à la page 38.)

[10] Abhinav Bhatele, Nikhil Jain, Yarden Livnat, Valerio Pascucci, and Peer-Timo Bremer. Analyzing network health and congestion in dragonfly-based supercomputers. In *Proceedings of Parallel and Distributed Processing Symposium*, pages 93–102. IEEE, 2016. (Cité aux pages 40 et 42.)

[11] Sourav S Bhowmick, Byron Choi, and Shuigeng Zhou. Vogue: Towards a visual interaction-aware graph query processing framework. In *Online Proceedings of the Conference on Innovative Data Systems Research*, 2013. (Cité à la page 14.)

[12] Bodo Billerbeck, Falk Scholer, Hugh E Williams, and Justin Zobel. Query expansion using associated queries. In *Proceedings of Information and knowledge management*, pages 2–9. ACM, 2003. (Cité à la page 14.)

[13] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 2005. (Cité à la page 47.)

[14] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. (Cité aux pages 32 et 58.)

[15] Ulrik Brandes and Christian Pich. An experimental study on distance-based graph drawing. In *Proceedings of International Symposium on Graph Drawing*, pages 218–229. Springer, 2008. (Cité à la page 47.)

[16] Lee Byron and Martin Wattenberg. Stacked Graphs - Geometry & Aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252, 2008. (Cité à la page 95.)

[17] Zhuang Cai, Yi-Na Li, Xianjun Sam Zheng, and Kang Zhang. Applying feature integration theory to glyph-based information visualization. In *Proceedings of Pacific Visualization Symposium*, pages 99–103. IEEE, 2015. (Cité à la page 4.)

[18] Yi Chen, Xinyue Zhang, Yuchao Feng, Jie Liang, and Hongqian Chen. Sunburst with ordered nodes based on hierarchical clustering: a visual analyzing method for associated hierarchical pesticide residue data. *Journal of Visualization*, 18(2):237–254, 2015. (Cité aux pages 40 et 41.)

[19] Zaida Chinchilla-Rodríguez, Benjamín Vargas-Quesada, Yusef Hassan-Montero, Antonio González-Molina, and Félix Moya-Anegóna. New approach

to the visualization of international scientific collaboration. *Information Visualization*, 9(4):277–287, 2010. (Cité aux pages 40 et 42.)

[20] Isaac Cho, Wewnen Dou, Derek Xiaoyu Wang, Eric Sauda, and William Ribarsky. Vairoma: A visual analytics system for making sense of places, times, and events in roman history. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):210–219, 2016. (Cité à la page 90.)

[21] Jihye Choi, Youngtae Cho, Eunyoung Shim, and Hyekyung Woo. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC Public Health*, 16(1), December 2016. (Cité à la page 68.)

[22] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of Conference on Human Factors in Computing Systems*, pages 443–452. ACM, 2012. (Cité aux pages 40 et 41.)

[23] Filip Claes, Dmitry Kuznetsov, Robin Liechti, Sophie Von Dobschuetz, Bao Dinh Truong, Anne Gleizes, Daniele Conversa, Alessandro Colonna, Ettore Demaio, Sabina Ramazzotto, et al. The empres-i genetic module: a novel tool linking epidemiological outbreak information and genetic characteristics of influenza viruses. *The Journal of Biological Databases and Curation*, 2014, 2014. (Cité à la page 68.)

[24] Filip Claes, Dmitry Kuznetsov, Robin Liechti, Sophie Von Dobschuetz, Bao Dinh Truong, Anne Gleizes, Daniele Conversa, Alessandro Colonna, Ettore Demaio, Sabina Ramazzotto, et al. The empres-i genetic module: a novel tool linking epidemiological outbreak information and genetic characteristics of influenza viruses. *Database*, 2014, 2014. (Cité aux pages 79 et 102.)

[25] Kristin A Cook and James J Thomas. *Illuminating the path: The research and development agenda for visual analytics*. IEEE, 2005. (Cité aux pages 5 et 68.)

[26] Olivier Corby, Rose Dieng-Kuntz, Fabien Gandon, and Catherine Faron-Zucker. Searching the semantic web: Approximate query processing based on ontologies. *IEEE Intelligent Systems*, 21(1):20–27, 2006. (Cité à la page 14.)

[27] José Cortiñas Abrahantes, Andrey Gogin, Jane Richardson, and Andrea Gervelmeyer. Epidemiological analyses on african swine fever in the baltic countries and poland. *EFSA Journal*, 15(3):1–73, 2017. (Cité à la page 68.)

[28] Jan De Leeuw and Patrick Mair. Multidimensional scaling using majorization: Smacof in r. *Department of Statistics*, 31(3):1–30, 2011. (Cité aux pages 56 et 58.)

[29] Mike DeBoer. Understanding the heat map. *Cartographic Perspectives*, (80):39–43, 2015. (Cité à la page 75.)

[30] Geoffrey M Draper and Richard F Riesenfeld. Interactive fan charts: A space-saving technique for genealogical graph exploration. In *Proceedings of Annual*

Workshop on Technology for Family History and Genealogical Research. Citeseer, 2008. (Cité aux pages 40 et 41.)

[31] Cody Dunne, Michael Muller, Nicola Perra, and Mauro Martino. Vorograph: Visualization tools for epidemic analysis. In *Proceedings of Human Factors in Computing Systems*, pages 255–258. ACM, 2015. (Cité aux pages 75 et 79.)

[32] Tim Dwyer, Kim Marriott, and Peter J Stuckey. Fast node overlap removal. In *Proceedings of International Symposium on Graph Drawing*, pages 153–164. Springer, 2005. (Cité aux pages 47, 62, 63, 64 et 65.)

[33] Tim Dwyer, Kim Marriott, and Peter J Stuckey. Fast node overlap removal—correction. In *Proceedings of International Symposium on Graph Drawing*, pages 446–447. Springer, 2006. (Cité à la page 64.)

[34] Mennatallah El-Assady, Valentin Gold, Carmela Acevedo, Christopher Collins, and Daniel Keim. Contovi: Multi-party conversation exploration using topic-space views. *Computer Graphics Forum*, 35(3):431–440, 2016. (Cité aux pages 40 et 42.)

[35] Samiha Fadloun, Pascal Poncelet, Julien Rabatel, Mathieu Roche, and Arnaud Sallaberry. Node overlap removal for 1d graph layout. In *Proceedings of Information Visualisation*, pages 224 – 229, 2017. (Cité aux pages 21, 47 et 106.)

[36] Samiha Fadloun, Arnaud Sallaberry, Alizé Mercier, Elena Arsevska, Pascal Poncelet, and Mathieu Roche. [demo] Integration of Text-and Web-Mining Results in EpidVis. In *Proceedings of International Conference on Applications of Natural Language to Information Systems*, pages 437–440. Springer, 2018. (Cité aux pages 10 et 106.)

[37] Nivan Ferreira, Jorge Poco, Huy T Vo, Juliana Freire, and Cláudio T Silva. Visual exploration of big spatio-temporal urban data: a study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, 2013. (Cité à la page 77.)

[38] Christiaan Fluit, Marta Sabou, and Frank Van Harmelen. Ontology-based information visualization: toward semantic web applications. In *Visualizing the semantic web*, pages 45–58. Springer, 2006. (Cité à la page 14.)

[39] Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein. Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150–157, 2008. (Cité à la page 68.)

[40] Johannes Fuchs, Petra Isenberg, Anastasia Bezerianos, and Daniel Keim. A systematic review of experimental studies on data glyphs. *IEEE Transactions on Visualization and Computer Graphics*, 23(7):1863–1879, 2017. (Cité à la page 4.)

[41] Johanna Fulda, Matthew Brehmel, and Tamara Munzner. Timelinecurator: Interactive authoring of visual timelines from unstructured text. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):300–309, 2016. (Cité à la page 71.)

[42] Emden R Gansner and Yifan Hu. Efficient node overlap removal using a proximity stress model. In *Proceedings of International Symposium on Graph Drawing*, pages 206–217. Springer, 2008. (Cité aux pages 47, 62 et 64.)

[43] Emden R Gansner, Yehuda Koren, and Stephen North. Graph drawing by stress majorization. In *Proceedings of International Symposium on Graph Drawing*, pages 239–250. Springer, 2004. (Cité à la page 47.)

[44] Rohan Goel, Samiha Fadloun, Sarah Valentin, Arnaud Sallaberry, Pascal Poncelet, and Mathieu Roche. EpidNews: An epidemiological news explorer for monitoring animal diseases. In *Proceedings of Visual Information Communication and Interaction*, to appear 2018. (Cité aux pages 68 et 106.)

[45] Lorraine Goeuriot, Wendy Chapman, Gareth JF Jones, Liadh Kelly, Johannes Leveling, and Sanna Salanterä. Building realistic potential patient queries for medical information retrieval evaluation. *Language Resources and Evaluation, workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, 2014. (Cité à la page 14.)

[46] Robert L Harris. *Information Graphics: A Comprehensive Inllustrated Reference*. Oxford University Press, 1999. (Cité à la page 58.)

[47] Kunihiko Hayashi, Michiko Inoue, Toshimitsu Masuzawa, and Hideo Fujiwara. A layout adjustment problem for disjoint rectangles preserving orthogonal order. *Systems and Computers in Japan*, 33(2):31–42, 2002. (Cité aux pages 47, 63 et 65.)

[48] Orland Hoeber, Michael Brooks, Daniel Schroeder, and Xue Dong Yang. Thehotmap.com: Enabling flexible interaction in next-generation web search interfaces. In *Proceedings of Web Intelligence and Intelligent Agent Technology*, volume 1, pages 730–734. IEEE, 2008. (Cité aux pages 15 et 25.)

[49] Orland Hoeber and Xue Dong Yang. Interactive web information retrieval using wordbars. In *Proceedings of Web Intelligence*, pages 875–882. IEEE, 2006. (Cité à la page 15.)

[50] Orland Hoeber and Xue Dong Yang. Hotmap: Supporting visual exploration of web search results. *Journal of Association for Information Science and Technology*, 60(1):90–110, 2009. (Cité à la page 15.)

[51] Orland Hoeber, Xue Dong Yang, and Yiyu Yao. Visiq: Supporting visual and interactive query refinement. *Web Intelligence and Agent Systems: An International Journal*, 5(3):311–329, 2007. (Cité aux pages 16, 17 et 24.)

[52] Xiaodi Huang, Wei Lai, ASM Sajeev, and Junbin Gao. A new algorithm for removing node overlapping in graph visualization. *Information Sciences*, 177(14):2821–2844, 2007. (Cité aux pages 47, 63 et 65.)

[53] Hideo Joho, Claire Coverson, Mark Sanderson, and Micheline Beaulieu. Hierarchical presentation of expansion terms. In *Proceedings of Applied Computing*, pages 645–649. ACM, 2002. (Cité à la page 17.)

[54] Tomihisa Kamada, Satoru Kawai, et al. An algorithm for drawing general undirected graphs. *Information processing letters*, 31(1):7–15, 1989. (Cité à la page 47.)

[55] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Gorg, Jorn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. *Information visualization*, 4950:154–176, 2008. (Cité à la page 5.)

[56] Donald Ervin Knuth. *The Stanford GraphBase: a platform for combinatorial computing*, volume 37. Addison-Wesley Reading, 1993. (Cité à la page 56.)

[57] Yehuda Koren and David Harel. A multi-scale algorithm for the linear arrangement problem. In *Proceedings of International Workshop on Graph-Theoretic Concepts in Computer Science*, pages 296–309. Springer, 2002. (Cité à la page 46.)

[58] Yehuda Koren and David Harel. Axis-by-axis stress minimization. In *Proceedings of International Symposium on Graph Drawing*, pages 450–459. Springer, 2003. (Cité à la page 47.)

[59] Yehuda Koren and David Harel. One-dimensional layout optimization, with applications to graph drawing by axis separation. *Computational Geometry*, 32(2):115–138, 2005. (Cité à la page 47.)

[60] Martin Krzywinski, Inanc Birol, Steven JM Jones, and Marco A Marra. Hive plots—rational approach to visualizing networks. *Briefings in bioinformatics*, 13(5):627–644, 2011. (Cité à la page 20.)

[61] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009. (Cité aux pages 40 et 41.)

[62] Chandan Kumar, Wilko Heuten, and Susanne Boll. Visualization support for multi-criteria decision making in geographic information retrieval. In *Proceedings of Availability, Reliability, and Security*, pages 363–375. Springer, 2013. (Cité à la page 73.)

[63] Tao Lei, Rui Cai, Jiang-Ming Yang, Yan Ke, Xiaodong Fan, and Lei Zhang. A pattern tree-based approach to learning url normalization rules. In *Proceedings of World wide web*, pages 611–620. ACM, 2010. (Cité à la page 38.)

[64] Wanchun Li, Peter Eades, and Nikola Nikolov. Using spring algorithms to remove node overlapping. In *Proceedings of Information visualisation*, pages 131–140. Australian Computer Society, 2005. (Cité aux pages 63, 64 et 65.)

[65] Rey-Long Liu and Yi-Chih Huang. Medical query generation by term–category correlation. *Information processing et management*, 47(1):68–79, 2011. (Cité à la page 14.)

[66] Yarden Livnat, James Agutter, Shaun Moon, and Stefano Foresti. Visual correlation for situational awareness. In *Proceedings of Information Visualization*, pages 95–102. IEEE, 2005. (Cité à la page 79.)

[67] Dongning Luo, Jing Yang, Milos Krstajic, William Ribarsky, and Daniel Keim. Eventriver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics*, 18(1):93–105, 2012. (Cité à la page 71.)

[68] Kim Marriott, Peter Stuckey, Vincent Tam, and Weiqing He. Removing node overlapping in graph layout using constrained optimization. *Constraints*, 8(2):143–171, 2003. (Cité aux pages 47, 63, 64 et 65.)

[69] Ali M Messenger, Amber N Barnes, and Gregory C Gray. Reverse zoonotic disease transmission (zooanthroponosis): a systematic review of seldom-documented human biological threats to animals. *PloS one*, 9(2):e89055, 2014. (Cité à la page 2.)

[70] Kazuo Misue, Peter Eades, Wei Lai, and Kozo Sugiyama. Layout adjustment and the mental map. *Journal of Visual Languages and Computing*, 6(2):183–210, 1995. (Cité aux pages 47, 63 et 65.)

[71] Davide Mottin, Francesco Bonchi, and Francesco Gullo. Graph query reformulation with diversity. In *Proceedings of Knowledge Discovery and Data Mining*, pages 825–834. ACM, 2015. (Cité à la page 17.)

[72] Richard O'Donnell, Alan Dix, and Linden J Ball. Exploring the pietree for representing numerical hierarchical data. In *People and Computers XX—Engage*, pages 239–254. Springer, 2007. (Cité à la page 96.)

[73] Alexandre Perrot, Romain Bourqui, Nicolas Hanusse, Frédéric Lalanne, and David Auber. Large interactive visualization of density functions on big data infrastructure. In *Proceedings of Large Data Analysis and Visualization*, pages 99–106. IEEE, 2015. (Cité à la page 75.)

[74] Jordi Petit. Experiments on the minimum linear arrangement problem. *Journal of Experimental Algorithmics*, 8:2–3, 2003. (Cité à la page 46.)

[75] Donna J Peuquet. It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of american Geographers*, 84(3):441–461, 1994. (Cité à la page 69.)

[76] Christian Pich. *Applications of Multidimensional Scaling to Graph Drawing*. PhD thesis, Universität Konstanz, 2009. (Cité à la page 47.)

[77] Robert Pienta, Fred Hohman, Acar Tamersoy, Alex Endert, Shamkant Navathe, Hanghang Tong, and Duen Horng Chau. Visual graph query construction and refinement. In *Proceedings of Management of Data*, pages 1587–1590. ACM, 2017. (Cité à la page 14.)

[78] Sébastien Rufiange and Michael J McGuffin. Diffani: Visualizing dynamic graphs with a hybrid of difference maps and animation. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2556–2565, 2013. (Cité aux pages 58 et 61.)

[79] David Sanchez, Laura Martinez-Sanahuja, and Montserrat Batet. Survey and evaluation of web search engine hit counts as research tools in computational linguistics. *Information Systems*, 73:50 – 60, 2018. (Cité à la page 38.)

[80] Purvi Saraiya, Peter Lee, and Chris North. Visualization of graphs with associated timeseries data. In *Proceedings of Information Visualization*, pages 225–232. IEEE, 2005. (Cité à la page 61.)

[81] Roeland Scheepens, Huub van de Wetering, and Jarke J. van Wijk. Non-overlapping aggregated multivariate glyphs for moving objects. In *Proceedings of Pacific Visualization Symposium*, pages 17–24, 2014. (Cité à la page 109.)

[82] Felix Schmitt, Robert Dietrich, René Kuß, Jens Doleschal, and Andreas Knüpfer. Visualization of performance data for mpi applications using circular hierarchies. In *Proceedings of Visual Performance Analysis*, pages 1–8. IEEE, 2014. (Cité aux pages 40 et 41.)

[83] Milton H Shimabukuro, Edilson F Flores, Maria Cristina F de Oliveira, and Haim Levkowitz. Coordinated views to assist exploration of spatio-temporal data: a case study. In *Proceedings of Coordinated and Multiple Views in Exploratory Visualization*, pages 107–117. IEEE, 2004. (Cité à la page 76.)

[84] Krzysztof Śmietanka, Grzegorz Woźniakowski, Edyta Kozak, Krzysztof Niemczuk, Magdalena Frączyk, Łukasz Bocian, Andrzej Kowalczyk, and Zygmunt Pejsak. African swine fever epidemic, poland, 2014–2015. *Emerging infectious diseases*, 22(7):1201–1207, 2016. (Cité à la page 35.)

[85] Anselm Spoerri. How visual query tools can support users searching the internet. In *Proceedings of Information Visualisation*, pages 329–334. IEEE, 2004. (Cité à la page 77.)

[86] Anselm Spoerri. Rankspiral: Toward enhancing search results visualizations. In *Proceedings of Information Visualisation*, pages 18–18. IEEE, 2004. (Cité à la page 77.)

[87] John Stasko and Eugene Zhang. Focus+ context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Proceedings of Information Visualization*, pages 57–65. IEEE, 2000. (Cité aux pages 40 et 42.)

[88] Kilian Stoffel, John D Davis, Gerald Rottman, Joel Saltz, James Dick, William Merz, and Robert Miller. A graphical tool for ad hoc query generation. In *Proceedings of American medical informatics association*, pages 503–507. AMIA, 1998. (Cité à la page 14.)

[89] Nenad Stojanovic. Information-need driven query refinement. *Web Intelligence and Agent Systems: An International Journal*, 3(3):155–169, 2005. (Cité à la page 17.)

[90] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional databases. *Communications of the ACM*, 51(11):75–84, 2008. (Cité à la page 14.)

[91] Edward R Tufte and Glenn M Schmieg. The visual display of quantitative information. *American Journal of Physics*, 53(11):1117–1118, 1985. (Cité aux pages 4 et 5.)

[92] Wouter Van den Broeck, Corrado Gioannini, Bruno Gonçalves, Marco Quaggiotto, Vittoria Colizza, and Alessandro Vespignani. The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BioMed Central infectious diseases*, 11(1):37, 2011. (Cité à la page 79.)

[93] Manasi Vartak, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. Seedb: automatically generating query visualizations. *Proceedings of Very Large Data Bases*, 7(13):1581–1584, 2014. (Cité à la page 14.)

[94] Lujin Wang, Joachim Giesen, Kevin T McDonnell, Peter Zolliker, and Klaus Mueller. Color design for illustrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1739–1754, 2008. (Cité à la page 4.)

[95] Matthew O Ward, Georges Grinstein, and Daniel Keim. *Interactive data visualization: foundations, techniques, and applications*. A K Peters, 2010. (Cité aux pages 4 et 5.)

[96] Colin Ware. *Information visualization: perception for design*. Elsevier, 2012. (Cité à la page 4.)

[97] Martin Wattenberg. Arc diagrams: Visualizing structure in strings. In *Proceedings of Information Visualization*, pages 110–116. IEEE, 2002. (Cité à la page 46.)

[98] Ji Soo Yi, Youn ah Kang, and John Stasko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007. (Cité à la page 76.)

[99] Peipei Yi, Byron Choi, Sourav S Bhowmick, and Jianliang Xu. Autog: a visual query autocompletion framework for graph databases. *The VLDB Journal*, 26(3):347–372, 2017. (Cité à la page 17.)

[100] Jiawei Zhang, Benjamin Ahlbrand, Abish Malik, Junghoon Chae, Zhiyu Min, Sungahn Ko, and David S Ebert. A visual analytics framework for microblog data analysis at multiple scales of aggregation. *Computer Graphics Forum*, 35(3):441–450, 2016. (Cité aux pages 40 et 41.)